IN SILICO STUDIES OF NUCLEIC ACID COMPLEXES WITH PROTEINS, AND THERAPEUTIC SMALL MOLECULES

A thesis submitted for the degree of
Doctor of Philosophy at University College London

Jarmila Husby (*née* Chladová)

Cancer Research UK Biomolecular Structure Group
Department of Pharmaceutical and Biological Chemistry
UCL School of Pharmacy

November 2012

This thesis describes research conducted in the UCL School of Pharmacy between January 2009 and July 2012 under the supervision of professor Stephen Neidle, professor David Thurston (until December 2011), and Dr. Gary Parkinson (since January 2012). I certify that the research described is original and that any parts of the work that have been conducted by collaboration are clearly indicated. I also certify that I have written all the text herein and have clearly indicated by suitable citation any part of this dissertation that has already appeared in publication.

Signature: _____ Date: _____

Part of the presented work has been disclosed in:

Husby, J., Todd, A.K., Haider, S.M., Zinzalla, G., Thurston, D.E. and Neidle, S. Molecular dynamics studies of the STAT3 homodimer:DNA complex: relationships between STAT3 mutations and protein-DNA recognition.
J Chem Inf Model. **2012**, 52(5), 1179-92.

Islam, B., D'Atri, V., Sgobba, M., Husby, J. and Haider, S. Computational methods for studying G-quadruplex nucleic acids. In *Guanine Quartets: Structure and Application*; ed. Spindler, L. and Fritzsche, W.; RSC Publishing Cambridge **2012**, 194-211.

Nkansah, E., Shah, R., Collie, G.W., Parkinson, G.N., Palmer, J., Rahman, K.M., Bui, T.T., Drake, A.F., Husby, J., Neidle, S., Zinzalla, G., Thurston, D.E. and Wilderspin, A.F. Observation of unphosphorylated STAT3 core protein binding to target dsDNA by PEMSA and X-ray crystallography.
(manuscript submitted for publication, **2012**)

Nasiri, H.R., Bell, N., McLuckie, K., Berndl, S., Husby, J., Abell, C., Neidle, S. and Balasubramanian, S. Targeting the c-MYC G-quadruplex using Fragment-Based design Induces c-MYC Repression in Cells.
(manuscript in preparation, **2012**)

ACKNOWLEDGEMENTS

I would like to warmly thank my supervisor Professor Stephen Neidle, whom I respect very much, for his professional guidance, kind support, and for keeping me "on track", yet providing me with lots of independence. I thank Professor David Thurston for giving me the opportunity to develop my skills in such an exciting, yet demanding field of science, and I extend my thanks also to Dr. Gary Parkinson.

I can not quantify the effect that Dr. Alan Todd has had on my projects, and who has patiently taught me that there is a solution to every (not only) computational challenge. Without his valued support, the outcome of my theoretical research would not be the same.

In silico research can be at times very challenging (and frustrating), but also fantastically inspiring and rewarding. I feel privileged to have met, worked with, and learnt from Dr. Jamie Platts, Dr. Shozeb Haider, Dr. Giovanna Zinzalla, and Dr. Arturo Robertazzi, who has shown me that my ever-since passion for music and for science can indeed co-exist in an "equilibrium". I feel very lucky to have worked in such a warm and friendly environment of the BMSG group, created by the past and present members, and in particular Gavin, Alan, Nancy, Caterina, Aaron, Stephan and Mekala, whom I sincerely thank for making my PhD years very enjoyable. A special thought belongs to Irene Dougherty, whom I can not thank enough for her kind help throughout the four years. I also acknowledge Cancer Research UK, and UCL School of Pharmacy that have provided me with the much appreciated funding for my PhD studies.

I am immensely grateful to my parents, František and Jarmila, who have always supported me, have had the faith in me, and have given me the "wings to fly".

The last paragraph belongs to my husband Aaron, who has always been here for me on every single step of this "journey". Words are not enough here, so I simply say, Thank you.

ABSTRACT

In silico approaches to nucleic acid targeted drug discovery have been used in order to study duplex DNA, in complexes with proteins as well as more unusual form of G-rich DNA folded into higher-order structures termed as G-quadruplexes, in complexes with therapeutic small molecules. The overall aim of this work has been to provide insight into the stability, recognition, energetics of binding and dynamic behavior of these DNAs in complexes with the STAT3βtc homodimer:DNA complex and with therapeutic small molecules in G-quadruplex/pyridostatin and G-quadruplex/fragment complexes by means of combined in silico approaches. The techniques of explicit solvent molecular dynamics (MD) simulations, and subsequent calculations of the free energies of binding, molecular docking, and 3D-pharmacophore modeling have been applied to study STAT3 and G-quadruplex DNA, promising targets for anticancer therapeutic intervention.

Analysis of the data obtained from multiple 50-ns MD simulations of the STAT3-DNA complexes has suggested how the transcription factor STAT3 interacts with duplex DNA, the nature of the conformational changes, and ways in which function may be affected. A majority of known pathologic mutations affecting the DNA-biding region of the STAT3 have been found at the protein-DNA interface, and they have been mapped in detail. The STAT3 conformations obtained from these MD simulations have been subsequently used as a basis for a comparative multiple-target molecular docking study with an in-house library of potential STAT3 inhibitors, providing a rational of their binding in the absence of structural data.

A novel “dynamic docking” approach (robust platform of numerous MD simulations) has been developed to address the G-quadruplex receptor and ligand flexibility issue, and subsequent conformational change upon binding. The strength of binding at different regions and both sites of the G-quadruplex were then closely examined. An *in silico* study of a fragment-based approach towards G-quadruplex stabilizing ligands has also been explored, in parallel with experimental studies, to assess whether this could provide a reliable rapid approach to finding hit fragments in the case of the c-MYC promoter quadruplex.

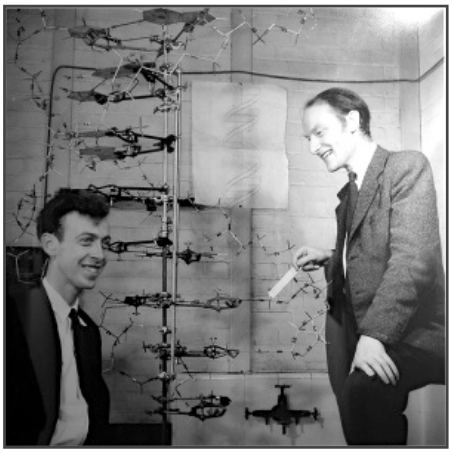
TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

LIST OF ABBREVIATIONS

Å	Angstrom (= 10^{-10} nm)
ANOVA	Analysis of Variance
bp	base pair
DNA	Deoxyribonucleic acid
*ds*DNA	double-stranded Deoxyribonucleic acid
FP	Fluorescent-Polarization
G4	Guanine-quartet
GA	Genetic Algorithm
GAFF	General AMBER Force Field
HB	Hydrogen Bond
HTS	High Throughput Screening
IC_{50}	Inhibitory concentration: concentration of a compound required to inhibit a biological process by 50%
IDA	Intercalator-Displacement-Assay
ITC	Isothermal Titration Calorimetry
FEP	Free Energy Perturbation
GAS	IFN-γ-activated sequence (site)
JAK	Janus Kinase
LIE	Linear Interaction Energy
MD	Molecular Dynamics
MM	Molecular Mechanics
MM/GBSA	Molecular Mechanics Generalized Born Surface Area
MM/PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MRC	Multiple Receptor Conformation
NMR	Nuclear Magnetic Resonance
NPT	isothermal-isobaric ensemble
NVE	microcanonical ensemble
NVT	canonical ensemble
PCA	Principal Component Analysis
PEMSA	Protein Electrophoretic Mobility Shift Assay
PDB	Protein Data Bank
PDI	protein-DNA interaction
PPI	protein-protein interaction
pY	phosphorylated tyrosine residue
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SAR	Structure-Activity Relationship
SASA	Solvent-accessible Surface Area
SH2	Src Homology 2
p-SH2	phosphorylated SH2 domain
u-SH2	unphosphorylated SH2 domain
SPE	Single Point Energy
STAT	Signal Transducer and Activator of Transcription
STAT3	Signal Transducer and Activator of Transcription 3
STAT3βtc	Signal Transducer and Activator of Transcription 3, isoform β, truncated
pSTAT3	phosphorylated STAT3
uSTAT3	unphosphorylated STAT3
TI	Thermodynamic Integration
TF	Transcription factor
TO	thiazole orange
VdW	van der Waals
VEGF	Vascular Endothelial Growth Factor
VS	Virtual Screening

INTRODUCTION



Watson and Crick with their DNA helix model in 1953
adapted from http://biologyinculture.wordpress.com

"DNA neither cares nor knows. DNA just is. And we all dance to its music."
(Richard Dawkins)

‘Overture’

The work described in this thesis is principally focused on *in silico* studies of nucleic acid complexes with proteins, and therapeutic small molecules. Whereas the fundamental role of proteins in biological processes has been recognized for many years (as most human diseases have a protein-related foundation), the importance of folded DNA within the framework of these processes is only more recently coming into prominence.

Despite being relatively young, both fields of molecular modeling (computational), and biomolecular structure determination (experimental) have undergone tremendous progress, since molecular mechanics developed around the same time that protein crystallography was coming to prominence, and the structure of the DNA was discovered by Watson, Crick, Franklin and Gosling.[1-3] The discovery of the elegant, yet simple DNA helix gave rise to molecular biology, and can be also seen as an imaginary cross-disciplinary connection, that has led to the completion of the human genome sequencing project in 2007. Additionally, *in silico* simulations have provided even more details of the structures of DNA and its complexes. Consequently, biomolecular simulations are now viewed as essential aspects of the modern drug discovery process.

So it follows that this work should be introduced by a concise literature review of the rich field of biomolecular modeling and its techniques, as well as the vast field of DNA as a molecular target for anticancer therapy. Only the most relevant aspects of the molecular targets of interest, and their “computation” will be discussed, describing the underlying principles of *in silico* approaches applied here.

CHAPTER 1:
DNA structure and recognition

Nucleic acids, in addition to forming the well-known DNA double helix, can form a variety of different (canonical and non-canonical) structures. Interestingly, the double helix is largely independent of its sequence, whereas the stability and form of other structures of DNA are often dictated by their sequence, and in particular by the different chemical properties of their nucleobases.[4]

1.1 STRUCTURAL DIVERSITY OF DNA AND ITS FUNDAMENTAL BUILDING BLOCKS

1.1.1 Essential components of DNA architecture

The elementary building block of nucleic acid polymers, the mononucleotide, is composed of a 5-membered, typically non-planar, sugar ring (deoxyribose in DNA and ribose in RNA), a phosphate group, and a purine (adenine and guanine) or pyrimidine (cytosine and thymine) base. Thymine is replaced by uracil in RNA. Nucleosides, the units formed from the sugar and base only, are then linked together via a phosphate group, forming a polynucleotide chain. Subsequently, (upon polymerization into DNA chains) a sugar/phosphate backbone is formed via a phosphodiester bridge joining the C3'-hydroxyl group of *n*-th nucleotide sugar with C5'-hydroxyl of the *(n+1)*-th nucleotide (3' to 5' phosphodiester bond) (Figure 1.1). In the classical DNA double helix, two opposite- (anti-parallel) direction (5'→ 3') polynucleotide chains form a flexible "ladder-like" structure with an imaginary central axis.

The conformational space of a nucleotide is specified by six backbone torsion angle rotations (α, β, γ, δ, ε, and ζ) and one glycosidic torsion angle χ, together with sugar pucker flexibility. There is a standard atomic numbering scheme in place for DNA. The DNA backbone, with the charged phosphate groups of the backbone exposed to the external environment (i.e. solvent), is specified by the following sequence: P → O5'→ C5'→ C4' → C3' → O3' (→ P) (Figure 1.1).

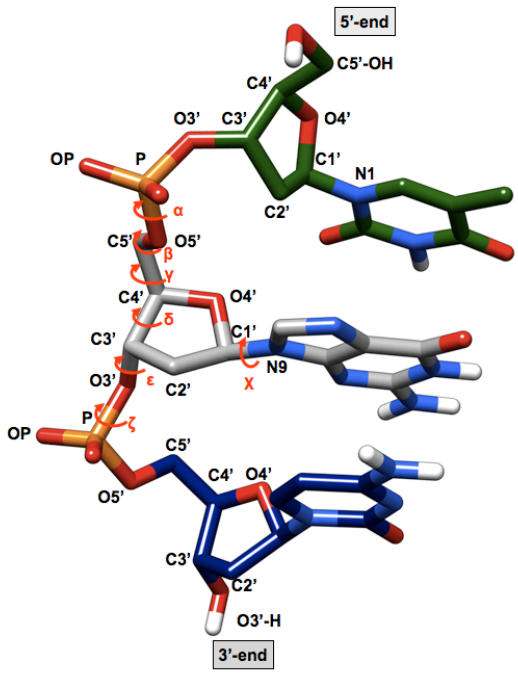


Figure 1.1: Structural features of a DNA chain with standard atom labeling.
The chain runs from the 5' (C5') to 3' (C3') end, with the nucleotide phosphodiester bonds linking the C3'-hydroxyl group of one nucleotide sugar with C5'-hydroxyl group of the next one. Nucleic acid torsion angles are labeled as α (O3'-P-O5'-C5'), β (P-O5'-C5'-C4'), γ (O5'-C5'-C4'-C3'), δ (C5'-C4'-C3'-O3'), ε (C4'-C3-O3'-P), ζ (C3'-O3'-P-O5') and χ (O4'-C1'-N1/N9-C2/C4 pyrimidine/purine). C2'-*endo* sugar pucker, and a distance between two consecutive P atoms ~ 6 Å are features typical for B-DNA.

Non-covalent interactions, in particular hydrogen bonding and stacking interactions, determine the structure of biomolecules[5,6] (such as nucleic acids and proteins). While it is understood that hydrogen bonding is essential for the specificity of base pairing, π-π stacking interactions between planar aromatic rings of nucleobases are equally important contributions to the final stability of nucleic acid structures.[6,7] Although individually weak, the additive power of these interactions has large cooperative stabilizing effects (π-π stacking ~ "deus ex machina").

1.1.2 Duplex DNA versus G-quadruplex DNA

The most prevalent form of DNA is the canonical anti-parallel double-helical duplex, linked together by Watson-Crick hydrogen bonding pairs; A:T (two hydrogen bonds), and G:C (three hydrogen bonds). The asymmetry in the base pairs gives rise to two parallel types of grooves (major and minor), whose dimensions reflect base pair (bp) distances from the central axis, and their orientation. There are a number of biologically-relevant helix forms (A-DNA, B-DNA or Z-DNA), however the right-handed B-form is dominant under physiological conditions[8,9] (Figure 1.2 a). B-DNA is 20 Å wide, and has 10 bp per turn, with the major groove 11.6 Å wide and 8.5 Å deep, while the minor groove is 6 Å wide and 8.2 Å deep. Both major- and in particular minor grooves have been extensively targeted with small-molecules for therapeutic intervention, with the aim of disrupting the transcription of specific genes at the DNA level.[10-13]

An alternative arrangement of DNA into "four-stranded" structures, called G-quadruplexes, is built from G-rich sequences of telomeric (and also genomic) DNA, around two or more π-π stacked quartets of hydrogen-bonded guanine bases (i.e G-quartets, G4), via Hoogsteen pairings, with an essential alkali metal ion (and an order of preference $K^+ > Na^+$) positioned in the interior channel of each G-quartet. The ions co-ordinate to O6 atoms of the guanine bases (Figure 1.2 b) resulting in a significant stabilizing effect to the structure. The G-quartets are linked together by intervening variable-length sequences, loops, arranged on the outside of the relatively rigid and stable G4 core. The flexible loops are thus the key determinants of G-quadruplex structural variability. Unlike duplex DNA, G-quadruplexes may adopt parallel, anti-parallel or mixed strand orientation, depending on the sequence, and experimental conditions.[14,15] There are three principal categories of quadruplex arrangements, that are possible: (1) tetramolecular, (2) bimolecular or (3) intramolecular (i.e unimolecular, where all guanine bases involved in G-tetrad formation originate from single strand of DNA). Depending on the G-quartet arrangement and the sequence length, the loops intervening between successive G-quartets can then adopt different conformations (lateral, propeller or diagonal).[16]

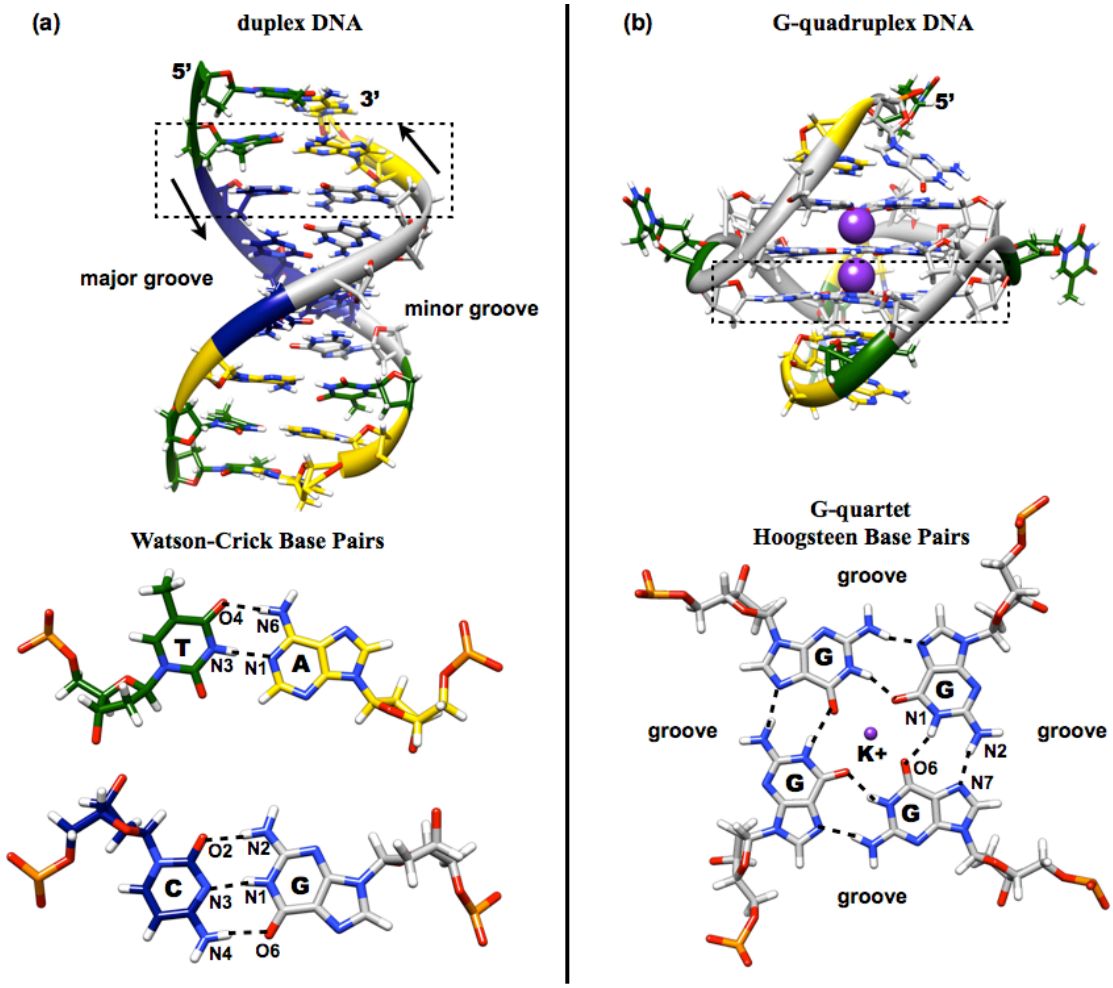


Figure 1.2: Essentials of duplex and G-quadruplex DNA.
(a) duplex DNA helix is formed by two anti-parallel right-handed strands, linked together by Watson-Crick base-pairing between A:T and C:G (i.e hydrogen bonding), together with π-π stacking interactions, that strongly contribute to the overall duplex stability. A base-pair is then the motif repeated 10-times within one helical turn. Two types of non-equivalent asymmetrical grooves are typical for duplex DNA;
(b) G-quadruplex DNA, despite being often described as "four-stranded", it can be formed from four (tetramolecular), two (bimolecular) or one (unimolecular) G-rich strand, organized into parallel, anti-parallel or mixed right-handed structures of π-π stacked G-quartets with an essential monovalent alkali ion located in the central channel. Hoogsteen base-pairing involves different face of the aromatic ring of nucleobases for hydrogen bonding. All G-quadruplex structures have four grooves, defined as the cavities bounded by the phosphodiester backbone.[16]

Conformation and properties of nucleic acids are directly influenced by solvation, with water molecules (well organized around bases and phosphates) being an integral part of nucleic acid structure.[17] DNA conformational type and base identity is reflected in hydration patterns (hydration sites), which represent sites of preferred polar binding to DNA. Calculated hydration sites can then be translated into predictions of amino acids binding to DNA in protein-DNA complexes.[18]

1.2 DNA IN COMPLEXES WITH PROTEINS

Interactions of nucleic acids with a wide range of DNA-binding proteins (such as regulatory, replication, recombination proteins etc.), are at the core of essential biological processes. There are a number of factors (structural features) that are of key importance in terms of nucleic acid recognition by other molecules; phosphate-phosphate distances in canonical DNA helices, minor/major groove dimensions which are directly related to their electron density (T:A pairs are electron poor compared to C:G), or stereochemical repulsion.[18] However, it is the specificity of these interactions that is fundamentally important in all aspects of gene regulation.[19,20] Thus large areas of research have evolved around protein-DNA complexes, with focus on their biomedical importance and systematic classification.[21]

1.2.1 Targeting transcription factors for therapeutic intervention

Transcription factors (TF) are DNA-binding proteins that, usually in combination with other proteins form the pre-initiation complex, regulate the synthesis of complementary RNA copies (messenger RNA, mRNA) of their target genes (i.e transcription) by controlling the recruitment of RNA polymerase II. Through subsequent translation into proteins, they determine the cellular phenotype. Interconnected signaling cascades are initiated upon binding their ligands (by cell surface receptors), involving mostly kinases and phosphatases that act on transcription factors, leading to activation of transcription machinery upon its recruitment.

Constitutive activation of transcription factors, however, can lead to various cancers. In particular three signaling pathways, namely the JAK/STAT pathway,[22] the Ras/Raf/MAP/ERK pathway[23] and the Hedgehog pathway,[24] have been found upregulated in a range of human cancers, making them of great interest for the development of new chemotherapeutic agents.[25] Molecularly-targeted therapies are fast becoming a real option for the treatment of cancers and other diseases. In malignant cancer growth, intracellular signaling pathways are dysregulated, which causes abnormal cell proliferation,

resistance to cell death (apoptosis), altered differentiation, and cytotoxic stress. Altering transcription factor activity can have significant consequences, since a multiple signaling pathway may converge on a single transcription factor, which can modulate a gene expression program that leads to oncogenic alteration.[26] Assessing and validating the druggability of molecular targets, and in particular transcription factors (they typically have both protein-protein and protein-DNA interfaces) has proven to be challenging. [27,28] However a number of transcription factors, such as STAT3 are viewed as suitable therapeutic - druggable - targets.[29]

1.2.2 Signal Transducer and Activator of Transcription 3 (STAT3)

Signal Transducers and Activators of Transcription (STATs) are a group of latent transcription factors in cytokine signaling, discovered in the early 1990's. These factors are unique due their ability to transduce signals from the cell membrane to the nucleus, where they activate gene transcription, bypassing secondary messengers.[30] Seven TF members of the STAT family have been identified, i.e. STAT1, 2, 3, 4, 5a, 5b and 6, and their activation is an essential event for the mediation of cytokines and growth-factor induced cellular and biological processes, such as proliferation, differentiation, survival, development and inflammation.[31,32]

STAT3 is the key mediator of interleukin 6 (IL-6)-type cytokine signaling.[33] Structurally, STAT3 is similar to other STAT proteins (closely related is STAT1), having the N-terminal domain involved in tetramerization,[34] while the coiled-coil domain plays a role in STAT3 activation.[35,36] The DNA-binding domain with sequence specificity for a palindromic IFN-γ-activated sequence (GAS) element[37] is also involved in nuclear translocation, and mediates co-regulatory interactions with other TF such as NF-κB.[38] The Src-homology 2 (SH2) domain is generally the most structurally conserved domain among STATs, and it is involved in receptor recruitment and also in STAT dimerization.[31] In contrast, the transactivation domain is the least preserved domain, and it is associated in PPI (for instance with CREB-binding protein).[39] The transactivation domain also accommodates Y705 and S727 (STAT3α only), the sites of tyrosine/serine

phosphorylation respectively, associated with STAT3 activation and dimerization.[40] Further post-translational modifications of STATs, such as lysine acetylation (K685),[41] or ubiquitination, modulate the transcriptional activity and significantly contribute to STAT-induced gene response.[31]

STAT3 exists in two isoforms of different properties[42]: the full-length STAT3α (~92 kDa) and the truncated STAT3β (~83 kDa), which lacks ~50 residues of the C-terminal transactivation domain. STAT3β is often considered as the dominant negative form, because its over-expression can suppress specific STAT3 functions[43,44] (Figure 1.3 a). However, STAT3β has been shown to activate specific STAT3 genes, demonstrating a unique function for the STAT3 isoforms.[45]

STAT3 is a well-recognised and important mediator of tumour-induced immunosuppression at many levels.[46,47] STAT3 is activated by numerous cytokine signaling pathways (i.e IL-6), and also by abnormal signalling of various growth factor receptors, together with oncoproteins such as Src and BCR-ABL.[48] As a key molecule in linking oncogenic pathways with immunosuppression, activated STAT3 not only down-regulates Th1 cytokines (critical mediators of anti-tumour immune response), but also directs malignant progression via the aberration of key proteins such as the cell survival (anti-apoptotic) proteins Bcl-x_L and Mcl-1, the cell-cycle regulators cyclin D1/D2 and c-MYC, and inducers of angiogenesis (i.e VEGF).[22,30] As a result, cancer cells with aberrantly active STAT3 show greater resistance to the initiation of apoptotic processes from the environment and apoptosis-initiating chemotherapeutics.[49] According to the original canonical view, multiple distinct steps are involved within the STAT3 signalling pathway (Figure 1.3 b). Upon extracellular receptor stimulation by growth factors or cytokines (e.g. IL-6), the receptor (*gp130*) becomes dimerized, and subsequently activated, which triggers the activation of the Janus protein tyrosine kinases JAK1 and JAK2, that are associated with the cytoplasmic tail of the receptor. This phosphorylation of the cytoplasmic tail provides a docking site for the recruitment of monomeric[30] or dimeric[50] un-activated STAT3 proteins via reciprocal interaction of their SH2 domains. Activated JAK tyrosine kinases phosphorylate recruited STAT3 proteins at a specific tyrosine within the C-terminus (Y705), and the phosphorylated STAT3 monomers dis-

sociate from the receptor to form STAT3-STAT3 homodimers. These then translocate from the cytoplasm to the nucleus, where they bind to target DNA motifs,[22] promoting the expression of proteins crucial for cell growth and survival. Normal STAT3 activation is relatively brief and regulated by a number of deactivation mechanisms, but human cancer cells express constitutively active STAT3 at high concentrations.[46] There is a mounting body of evidence[51] suggesting that tumour formation can be caused by abnormally active STAT3. Thus inhibition of aberrant STAT3 activity induces growth arrest and apoptosis of tumour cells *in vitro* as well as *in vivo*, validating STAT3 as a suitable molecular target for anticancer drug discovery.[22,51,52] Due to the complexity of the signalling pathway (as described above), targeting STAT3 in order to suppress is aberrant function in cancer cells can be approached (1) indirectly, by targeting the upstream components of the STAT3 pathway, or (2) directly targeting the STAT3 protein by means of SH2 domain inhibitors (dimerization inhibitors), the DNA-binding domain inhibitors or the N-terminal domain inhibitors.[51] Inhibitors targeting the STAT3 SH2 domain are the most explored and reported on strategy.

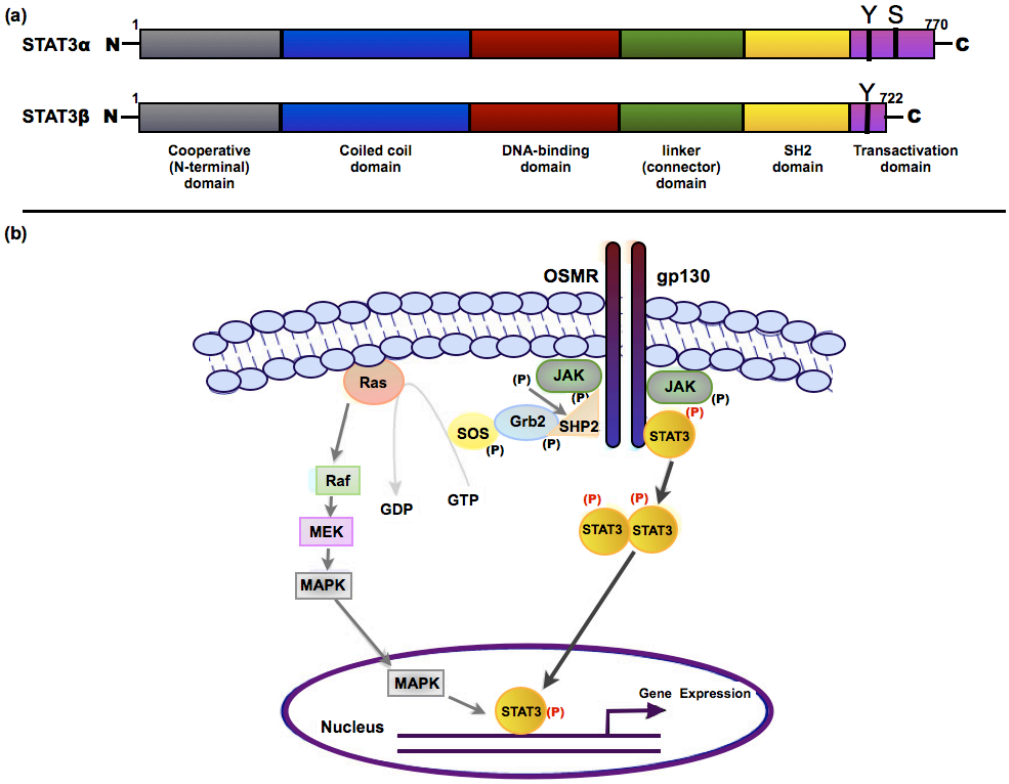


Figure 1.3: Functional domains of STAT3 protein; and canonical STAT3 signalling pathway.
(a) the full length STAT3α, and its isoform STAT3β; (b) Canonical view of STAT3 signalling cascade;

1.2.3 STAT3 inhibitors: from peptides to small molecules

For the first time, STAT3 dimerization was directly inhibited by Turkson *et al,* employing a truncated peptide sequence derived from the native STAT3-binding phosphopeptide (PpYLKTK).[53] A number of research groups have since developed (and patented) short peptides[54,55] and peptidomimetics[56-58] designed to block the SH2 domain of STAT3. However, membrane permeability and stability are the two main challenges facing a peptidic approach to drug discovery.[51] Other classes of therapeutic molecules designed to inhibit STAT3 function are oligonucleotides[59,60] (and even G-quartet oligonucleotides[61]), platinum-based compounds[62] and a large field of small molecule-inhibitors, mostly discovered via *in silico* screening an/or rational drug design.[49] However, only five classes of small molecule inhibitors,[63-67] and their second generation compounds,[68-72] have been patented so far (summarized in Figure 1.4) with only one inhibitor, BP-1-102, being reported as orally bioavailable.[72]

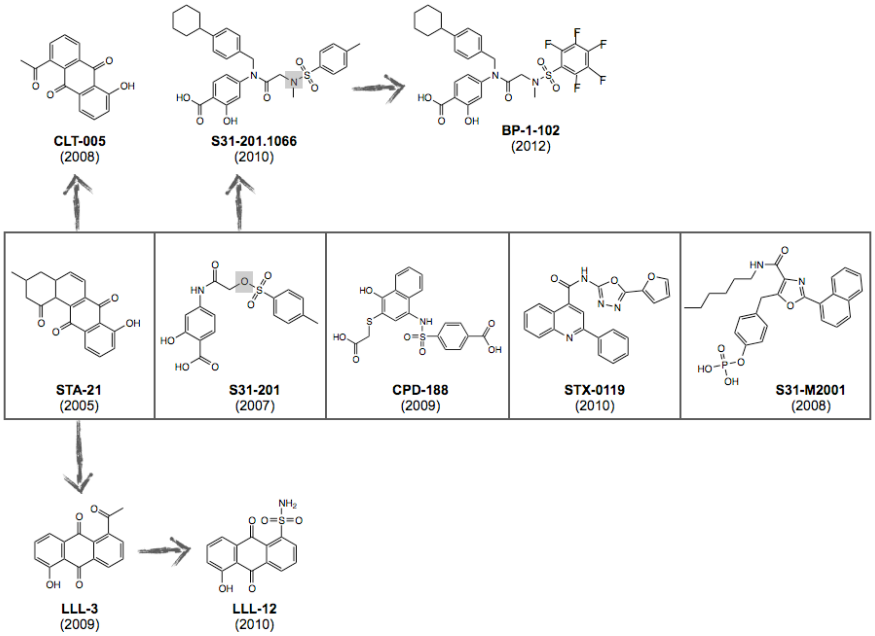


Figure 1.4: Patented small molecule inhibitors of STAT3 and their analogues, discovered via *in silico* approaches and/or rational structure based drug design.
(1) natural product STA-21[63] identified by VS (DOCK) of 429.000 compounds at the NCI, and its related compounds CLT-005[68] (patented for retinal disease treatment), LLL-3[69] and LLL-12[70]; (2) salicylic acid-based inhibitor S31-201[64] identified by VS (GLIDE) optimized to leading dimerization-disruptive agent S31-201.1066,[71] that was further optimized into BP-1-102,[72] a first orally bioavailable STAT3 inhibitor; (3) CPD-188[65] identified by in silico screen (~1.5 million compounds, MolSoft) employing complementary screens against STAT1; (4) STX-0119[66] identified by employing 3 different scoring techniques; (5) rationally-designed substituted oxazole S31-M2001,[67] based on structure of peptidomimetic ISS610.[56]

1.3 G-QUADRUPLEX DNA AS A THERAPEUTIC TARGET

G-quadruplex DNAs are involved in a wide range of biological and biochemical processes and have been located throughout the human genome.[73] Furthermore, they may have a significant promise as pharmacological targets in anticancer therapy. The therapeutic strategy for targeting G-quadruplexes initially emerged for telomeric DNA (and telomerase inhibition),[74,75] and within the past decade, interest in therapeutic significance of G-quadruplexes has increased, and now targets G-quadruplexes in gene promoters (i.e c-KIT, c-MYC, b-RAF).[76] Over the past decade, the conformational variance of G-quadruplex nucleic acid structures (both DNA and RNA) has been effectively studied by NMR and X-ray crystallography, and in particular in the last two years, nearly a third of the currently known quadruplex structures have been solved. To this date, there are over 150 quadruplex structures deposited in the Protein Data Bank (www.rcsb.org/pdb), both native and in complexes with ligands (i.e BRACO-19, [77] the porphyrin TMPyP4,[78] or several tetrasubsituted naphthalene diimide compounds[79]). The structural data provide a solid basis for rational drug design of new G-quadruplex interacting compounds, whose presence has been numerously shown to cause rearrangements of the bases within the loops, forming new ligand binding scaffolds.[79]

1.3.1 G-quadruplexes in human telomeres

Telomeres are nucleoprotein complexes located at the terminal regions of eukaryotic chromosomes, and in mammals, they are composed of non-coding tandem repeats of the sequence d(TTAGGG), together with associated telomeric proteins (six-protein complex) known as shelterin.[80] Telomeres play a major role in protecting telomeric region from degradation and genomic instability.[81] The terminal 100-200 nucleotides at the 3'-end of telomeres are single-stranded (overhang), and able to form G-quadruplex structures. Stabilization of G-quadruplexes in human telomeric DNA with small molecules is a promising anticancer strategy, as that has been shown to inhibit the activity of telomerase, which is over-expressed in ~ 80-85% of cancer cells and primary tumors, acting as a tumor promoter.[82]

1.3.2 Computational methods employed in studying G-quadruplex/ligand complexes

A number of computational studies, carried out to date, have employed biomolecular simulation methods to obtain a superior insight into the structure and interactions of both telomeric and genomic G-quadruplex nucleic acids. Dynamic behavior,[83,84] conformational properties[85] and energetics,[86,87] have been addressed, as well as plausible higher order structures,[88,89] cation binding,[90-92] or the behavior of existing G-quadruplex ligands.[93,94] A growing number of complementary molecular docking and ligand-based pharmacophore modeling studies have been regularly employed in parallel to experimental biophysical assays, in order explore plausible binding modes of the studied G-quadruplex ligands, to optimize the lead compounds, and to rationalize their selectivity. For instance, selectivity of a group of naphthalene diimide ligands for telomeric G4-RNA over the G4-DNA was explained by Collie[95] *et al* using the docking method in the AFFINITY program (Accelrys; http://accelrys.com/).

With respect to ligand-based pharmacophore modeling, two quadruplex-stabilizing alkaloid compounds, obtained by screening a Chinese herbal medicine compounds database (10,000 compounds) with previously generated pharmacophores (Catalyst software, Accelrys) have been described;[96] a ligand-based pharmacophore model, devised on the basis of acridine derivatives, was employed in the identification of triaryl-substituted imidazole derivative TSIZ01.[97] However, obtaining the most favourable binding poses of the compounds *in silico* has been challenging due to the specific features of quadruplex nucleic acid molecules. The highly charged backbones, the presence of stabilizing alkali metal cations, the basic quadruplex architecture and in particular the flexibility of the G-quadruplex structures are all important issues to be considered within the framework of *in silico* screening of G-quadruplex ligands by molecular docking. The quality of G-quadruplex docking results, and subsequently the binding affinity of potential G-quadruplex ligands, might be strongly affected by the flexibility of the loop regions.[98] To address the receptor and ligand flexibility issue, and subsequent conformational change upon binding, a novel form of fully “dynamic docking” method has been developed (described in CHAPTER 6).

Consequently, virtual screening of large chemical libraries to aid the selection of new G-quadruplex-binding ligands has not yet been extensively explored and, until recently, rarely used.[99] Two examples of G-quadruplex binding ligands are shown in Figure 1.5, both using the human telomeric quadruplex structure (PDB id 1KF1) as a starting point, and employing the ICM method of molecular docking software (Molsoft[100]); a synthetic substituted indole[101] identified by *in silico* screening of ~100,000 drug-like compound, and the naphthopyrone natural product fonsecin B[102] identified by screening over 20,000 compounds in natural product database. The use of molecular docking in virtual screening for the identification of bioactive molecules from natural product databases has recently been reviewed by Ma[103] *et al*.

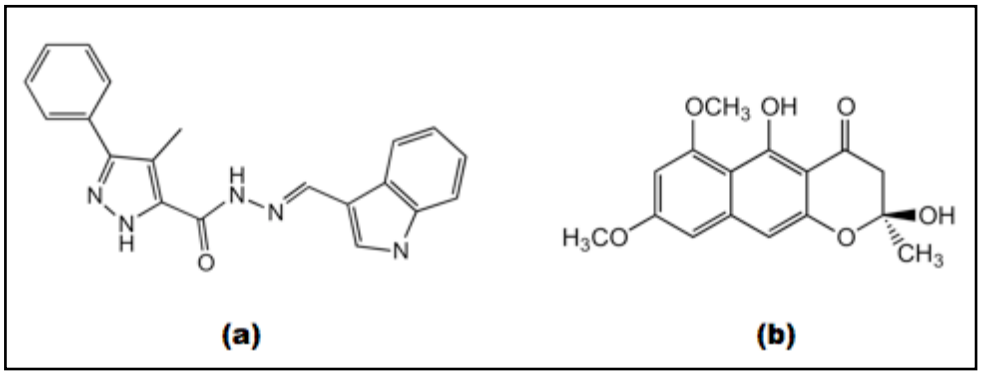


Figure 1.5: Two G-quadruplex-binding compounds identified by virtual screening.
(a) a synthetic substituted indole[101] ; (b) the naphthopyrone natural product fonsecin B[102]

In silico studies combined with NMR experiments in tandem provide more detailed understanding of the interactions between the ligands and G4-DNA.[104] A relatively small, but structurally very diverse commercially available database (6,000 compounds) was screened against the parallel structure of the simple quadruplex [d(TGGGGT)$_4$] (PDB id 1S45) by means of AutoDock v. 4. Subsequent NMR screening of the top 30 hits identified six G4-groove-binding molecules, that were further studied in detail by NMR, Isothermal Titration Calorimetry (ITC) measurements and molecular docking with modified quadruplexes, identifying the most potent G-quadruplex groove-binders to date.[105]

Truly large-scale integrated *in silico* and in vitro screening platforms for the purpose of discovering novel small molecules binding to specific nucleic acids were described by the Chaires' group, using the analogy of a "funnel".[106] The ZINC "drug-like" virtual database of 11.3 million compounds was screened against the antiparallel quadruplex target (PDB id 2HY9) employing the Surflex-Dock molecular docking software. An array of potential nucleic acids competing sites, as well as all possible binding sites on the target nucleic acid itself were considered for the virtual screening, which was performed on a grid of more than 10,000 computer processors. This approach, as well as the appropriate choice of the software, were previously validated.[107,108] The top 160 hits that emerged after scoring, for selective binding to the quadruplex target, were tested by a high-throughput melting assay followed by a secondary screening of the top compounds. A characterization of their binding behavior with the quadruplex target, by means of rigorous binding studies using calorimetry, spectroscopy, competition dialysis, molecular dynamics simulations and functional assays, identified a substantially stabilizing quadruplex-binder, which in turn suggests that the proposed in silico and in vitro platform may be used to discover new G-quadruplex binding scaffolds.[106]

At present, virtual screening alone is unable to adequately predict the selectivity of a particular ligand for different G-quadruplexes, because of the inherent inaccuracies in comparing binding energies between multiple G-quadruplex structures.[98] However G-quadruplex structures display a high degree of diversity in their loop and grooves geometries, which in particular could be used to enhance the selectivity of ligands, allowing for their excellent structure-specific recognition, affinity and specificity. For instance, in the case of the promoter KIT1 quadruplex structure (the NMR structure[109] or the X-ray structure[110]) the presence of a distinct cleft may be suitable for ligand binding and virtual HTS.[76]

CHAPTER 2:
Molecular modeling and computational approaches to biomolecular structure and drug design

'Molecular modeling is the science and art of studying molecular structure and function through model building and computation'.[111] The "model" is defined by the *Oxford English Dictionary* as 'a simplified and idealized description of a system or process, often in mathematical terms, devised to facilitate calculations and predictions'. The computations then include *ab initio* and semi-empirical quantum mechanics, empirical (molecular) mechanics, molecular dynamics, Monte Carlo, free energy and solvation methods, structure/activity relationships (SAR), chemical/biochemical information and databases, and a number of other established procedures. Furthermore, the refinement of experimental data, such as from X-ray crystallography or nuclear magnetic resonance (NMR), is also regarded as part of biomolecular modeling.

'The key in modeling is to develop and apply models that are appropriate for the questions being examined with them'.[111]

2.1 THE CONCEPTS AND PRINCIPLES OF MOLECULAR MODELING

Since starting in the 1960s, and progressing rapidly since the 1980s, together with the ever-increasing access to supercomputers the burgeoning field of biomolecular modeling has been in continuous development. Many ongoing methodological and technological improvements have provided a robust platform for *in silico* biomolecular studies, and computational chemistry/biology. Among those of key importance are: advances in instrumental and experimental techniques; novel models and improved algorithms for molecular simulations; and the availability of increasingly fast supercomputers, parallel processors and GPU computing.[112] However, communication, both

at the personal (i.e multidisciplinary collaborations between theoreticians and experimentalists), as well as virtual level (fast ubiquitously present Internet and web resources), play a pivotal role in facing the challenges of biomolecular modeling and *in silico* drug design.

Molecular modeling may also be classified as a division of science focused on applying the fundamental laws of physics and chemistry to the study of molecules, and biological macromolecules. Currently, it is one of the most-widely applied techniques in biological chemistry, biophysics and drug discovery.[113] The ultimate aspect enabling the widespread and interdisciplinary application of molecular modeling is the continuing growth of computing power. This has made it possible to analyze, compare and characterize large and complex data sets obtained from biomolecular systems experiments.[114] In the case of drug discovery, the principal aim is to create models and simulations, which can assist in the individual stages of a discovery pipeline by predicting, rationalizing and estimating the properties of molecules and their interactions, leading to more rational approach to the drug discovery and development process.[115] Rational drug design is then based on a fundamental assumption that drug activity is obtained via molecular binding of a ligand molecule to the pocket of a macromolecular target. Their chemical and geometrical complementarity in their active state is then essential for successful drug activity.[116]

The importance of molecular modeling and simulation methods has grown in the past decade due to their exquisitely detailed contributions to the study and function of biomacromolecules; e.g protein folding and conformational changes, modeling the dynamics of ion channels and transport across membranes, modeling and analysis of enzyme mechanisms, computation of binding free energies for ligands, structure-based drug design etc.[117] However, experimental data plays an key role in biomacromolecular modeling. For instance, development of the classical molecular mechanics force fields would not have been possible with such data, since quantum-chemical theoretical data alone would not be sufficient to develop a force field.[114]

2.2 MOLECULAR MECHANICS: A FOUNDATION FOR FORCE FIELDS

The notion that molecular geometry, energy, and various molecular properties may be calculated from mechanical-like models subjected to basic physics forces, underlines the basis of molecular modeling. A molecule, represented as a mechanical system in which the *particles* are connected by *springs*, rotates, translates and vibrates in order to adopt favourable conformations in space upon inter- and intra-molecular forces acting on it. The forces are depicted as a sum of harmonic-like (from Hooke's law) terms for bond-length and bond-angle deviations from equilibrium (ideal) values; trigonometric dihedral (torsional) terms to account for molecule's internal rotation; and non-bonded van der Waals and electrostatic potentials.[111]

2.2.1 Underlying principles of molecular mechanics

The term molecular mechanics (MM) refers to the use of simple potential energy functions (i.e. harmonic oscillators or Coulombic potentials) to model molecular systems.[118] As an alternative simplified approach to quantum mechanics (QM), the MM method ignores the electronic motions (electrons are treated implicitly) and calculates the energy of a system as a function of the nuclear position only, with the assumption of the Born-Oppenheimer approximation to the potential energy.

There are three underlying principles, originating from quantum-mechanical roots (i.e the Schrödinger equation), which determine the overall effectiveness of MM:

• The thermodynamic hypothesis, which assumes that a strong thermodynamic force drives "scrambled" conformations of high free energy to the native state of low free energy.

• The principle of additivity, assuming that the effective molecular energy can be expressed as a sum of potentials derived from simple physical forces (i.e bonded and non-bonded terms)

• The principle of transferability, which assumes that potentials can be developed to include all experimental data for *representative* structures, and subsequently be successfully applied to predict large biomolecules formed of adequate chemical subgroups.

MM approaches are generally applied in energy minimization, dynamic simulations (i.e molecular dynamics or Monte Carlo simulations), or ligand docking/scoring simulations, hence both effective formulation of the potential energy function as well as the suitable search algorithm are key aspect of a successful MM simulation.

2.2.2 Force field: functional form of potential energy function

'An empirical force field is a recipe for reproducing a potential energy surface'.[119] Potential energy is a component of the total energy of the system based on the position of the atoms (usually expressed in terms of Cartesian coordinates). Atoms are approximated as Lennard-Jones spheres with constant point charges localized within the atomic centers,[120] while force is the negative derivative of the potential, with respect to position. The equation for potential energy is known as the functional form of a force field. Each force field utilizes a slightly different functional form, however the equation always comprises bonded (bonds, angles, dihedrals, and improper dihedrals) and non-bonded (Lennard-Jones and Coulombic) terms (Figure 2.1, equation 2.1, 2.2 and 2.3).

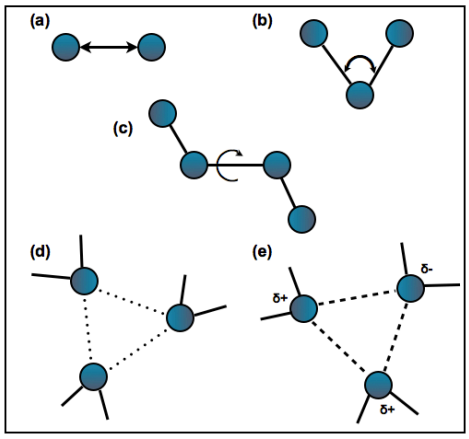


Figure 2.1: Representation of the key components to a molecular mechanics force field.
(a) bond stretching, (b) angle bending, (c) bond rotation (torsion), (d) non-bonded van der Waals interactions and (e) non-bonded electrostatic interactions.

$$E_{total} = E_{bonded} + E_{nonbonded} \quad (2.1)$$

$$E_{total} = \sum_{bonds} E_i^{bonds}(b_i) + \sum_{angles} E_i^{bond\ angles}(\theta_i) + \sum_{dihedrals} E_i^{tor}(\varphi_i) + \sum_{nonbonded\ pairs} E_{i<j}{}^{r}(r_{ij}) \quad (2.2)$$

The total energy (E) of a molecule is then defined as a sum of contributions arising from "ideal" bond length and bond angle distortions (b_i, θ_i) and internal torsion flexibility (φ_i) along with contributions from non-bonded (van der Waals and electrostatic) interactions (r_{ij}). One very typical and widely applied force field is AMBER (an acronym for Assisted Model Building with Energy Refinement). Like other widely applied force fields in consists of several discrete terms in simple functional form, describing an intra- and inter-molecular force within the system as follows (equation 2.3):

$$V(r) = \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 + \sum_{torsions} K_\phi[1+\cos(n\phi - \gamma)] + \sum_{\substack{nonbonded \\ pairs}} \left[\frac{q_i q_j}{\varepsilon r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}\right] \quad (2.3)$$

The first two terms of the above equation are harmonic potentials representing the interaction between two (bonds) and three (angles) atoms, separated by one (bonds) or two (angles) covalent bonds respectively. They are an approximation to the energy of a bond/angle, as a function of displacement from the reference (ideal) bond length b_0 or bond angle θ_0 respectively. The force constants K_b and K_θ then determine the strength of the bond/angle. The third term is periodic (i.e cosine function) torsion angle potential function, which models the presence of steric barriers between four atoms separated by three covalent bonds. The associated rotational motion is described in Fourier terms, where each rotational sequence described by a torsion (dihedral) angle ϕ, multiplicity n (i.e coefficients of symmetry), which denote the periodicity of the rotational barrier, and force constant K_ϕ which is the associated barrier height (magnitude). A reference phase angle (phase shift) γ denotes where the torsion angle passes through its minimum value. Additionally, improper torsion term potentials are used to describe out-of-plane bending frequencies, i.e to keep four atoms properly planar (hence the atoms involved are not

serially bonded but rather branched). Improper dihedrals can be expressed in terms of harmonic potentials, but in case of AMBER force field, they are reproduced in corresponding way as proper dihedrals, with the phase angle γ always equal to 180 degrees (π radians). Collectively, these three terms represent the internal or intramolecular parameters.

The non-bonded interactions (the fourth term of equation 2.3) contain a repulsion term and a dispersion term, describing the interactions (balance) between neutral atoms; and a Coulomb term, dealing with the electrostatic interactions. The repulsive and attractive terms (i.e. van der Waals interactions) are combined in the Lennard-Jones (12-6) potential, where the attractive and repulsive coefficients C and A depend on the chemical nature of the two interacting atoms (they are atom-pair specific). The attractive/repulsive effects become significant with decreasing inter-atomic distance r, thus positioning the atoms at the optimal distance stabilizes the system, giving rise to a minimum in the energy (Figure 2.2). The measure of how strongly two non-bonded particles attract each other (i.e the depth of the potential energy well) is referred to as 'epsilon' (ε), while 'sigma' (σ) is referred to as the van der Waals (atomic) radius.

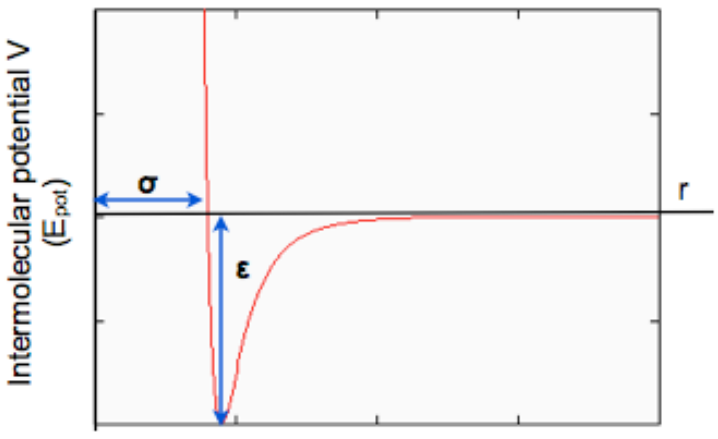


Figure 2.2: A graph of Lennard-Jones 12-6 potential.
Lennard-Jones 12-6 potential describes the attraction and repulsion forces between nonionic particles. ε and σ characterize the strength of the non-bonded interaction, and the atomic radius respectively. These atom pair parameters are then used to calculate the C and A attractive/repulsive coefficients.

Lastly, the electrostatic interaction between fully or partially charged groups (atoms) is represented by a Coulombic potential term, where two atoms (i and j) have point charges q_i and q_j. The magnitude of the electrostatic energy varies inversely with the distance between the atoms r_{ij}. The dielectric constant ε is typically set to 1, corresponding to permitivity in a vacuum (in explicit solvent simulations).

As mentioned earlier, the total energy of the system is calculated as the sum of all bonded and non-bonded interactions. However, the direct evaluation of the long-range electrostatic interactions would be computationally very complex, time consuming and expensive, therefore the non-bonded intermolecular distances are typically truncated at a cut-off distance (any interaction past the cutoff distance is ignored), which could potentially lead to an accumulation of errors. Use of the Particle Mesh Ewald (PME)[121,122] treatment of electrostatics, a technique that calculates the electrostatic energy of a system on a lattice with periodic boundary condition, can mostly eliminate the problem.

Electrostatics effects are long-range, and their influence on polar and charged molecules (water, ions, amino and nucleic acids) makes them of a special importance in terms of determining the structure, motion and function of many biomolecular systems (such as protein-DNA binding, biomolecule-ligand binding), as they are ubiquitous. Electrostatic interactions are also essential in determining the thermodynamics of binding[123] (i.e binding affinity), discussed later in the text.

Theoretical methods based on empirical force fields are often applied to biomolecular systems (i.e nucleic acids and their complexes) in the context of molecular dynamics (MD) simulations, that can further provide their atomic details, as well as free energies related to conformational changes of the studied biomolecular system.

2.3 MOLECULAR DYNAMICS SIMULATIONS OF BIOMOLECULES

In molecular dynamics (MD) simulations, the motion of a biomolecular system under the effect of a "force" (i.e a specified force field) is simulated by following its molecular configurations in time, according to Newton's equation of motion (the second law). In terms of individual atoms (particles), mathematically, the net force is equal to the time derivative (dt) of the particle's momentum (p), which is defined by its mass (m) and velocity (v), as shown in equation 2.4:

$$F = ma = m\,dv/dt = d(mv)/dt = dp/dt \tag{2.4}$$

However for a large biomolecular system of N atoms, the Newtonian equation of motion is written in terms of first-order differential equations (equation 2.5):

$$MV(t) = F(X) = -\nabla E(X(t)) + \ldots,$$
$$\dot{X}(t) = V(t) \tag{2.5}$$

where $X \in R^{3N}$ indicates the collective Cartesian vector of the system (i.e the x, y and z components of each atom); V is the corresponding collective velocity vector; M is the diagonal mass matrix, and the dot superscript indicates differentiation with respect to t. The total force (F) comprises the systematic force, which is the negative gradient of the potential energy (E) and, possibly, additional terms that mimic the environment.[124] Each gradient component i, i = 1,... 3N, is then given by equation 2.6:

$$\nabla E(X)_i = \partial E(X)/\partial \alpha_i, \tag{2.6}$$

where α_i indicates the x, y or z component of the atom. These equations must be then integrated numerically (by means of MD integrators such as Verlet or the leap-frog algorithm), providing a sequence of solutions and velocity pairs, $\{X^n, V^n\}$, for integers n that represent discrete times $t = n\Delta t$ at intervals (i.e timesteps) Δt. The initial atomic velocities are generated with a Maxwell-Boltzman distribution at given temperature.[124] Thus in summary, the resulting MD trajectories are then defined by both position and velocity vectors and they describe the evolution of the system in phase space.

As a consequence of Newton's law application, the total energy of the system is conserved, thus MD simulations naturally form a micro-canonical ensemble (NVE). Algorithms that introduce a thermostat or barostat to the system allow the sampling of the canonical (constant NVT) or isothermal-isobaric (conserved NPT) ensemble. MD follows the time evolution of a system, enabling us to study its dynamic properties, as all the degrees of freedom of the system are subjected to a force, hence move. It is often necessary to constrain some degrees of freedom using constraints algorithms to bonds such as the SHAKE[125] or LINCS[126] method. The latter method (LINCS, an acronym for Linear Constraint Solver) was employed in the molecular dynamics simulations described in this thesis.

MD simulations applied to biomolecular systems are a powerful technique that can provide us with valuable insight into important features of studied systems, where experimental data is not accessible. However, it is necessary to carefully consider their current limitations (i.e the sampling algorithms and time-scale limitations, force field approximation and polarization effects) of simulations in order to avoid over-interpretation of the results.

2.4 FREE ENERGY CALCULATIONS AS POST-PROCESSING METHODS

Biomolecular association events can be predicted by thermodynamics, i.e the extent of biochemical reactions, as well as the direction of spontaneous processes and stability. "Thermodynamics quantifies equilibrium, phase changes and stability using unmeasurable quantities like enthalpy and entropy; these are coupled to experimentally measurable quantities like temperature and pressure, through mathematical relationships".[127] In such a way, physicochemical transformations in micro- and macromolecular systems can be explained.

2.4.1 The basic concept of "free-energy"

The perception of "free energy" as a key criterion behind many important thermodynamic phenomena (such as equilibrium in chemical reactions), has been extensively studied for several decades now. Free energy is essentially "the factor that determines how a process will proceed and the probability that a system will adopt given state".[128] In terms of theoretical predictions, the expression for the free energy (Helmholtz free energy) in canonical ensemble (i.e NVT is constant) is directly related to its partition function, and is given by (equation 2.7):

$$F = -k_B T \ln Q_{NVT} \qquad \text{where} \qquad Q = \sum_i e^{-\beta E_i} \tag{2.7}$$

where k_B is Boltzmann's constant and Q_{NVT} is the partition function of the system (i.e describing its statistical properties in thermodynamic equilibrium), and β is the inverse temperature divided by Boltzmann's constant. The absolute free energy (equation 2.7) can only by directly calculated for small simple systems governed by a simple Hamiltonian, where an analytical expression for the partition function is obtainable.[128]

For most biomolecular systems, the aim is to determine relative free energies; for instance, the difference in binding of two ligands with the same receptor, or a complex formation (bound/unbound state). Because free energy is a state function, free

energy differences between distinct sub-states can indeed be evaluated without simulating the transition. Thus free energy differences between two states (A and B) are the ratios of partition functions (equation 2.8):

$$\Delta F = F_A - F_B = -k_B T \ln(Q_A/Q_B) \quad (2.8)$$

The ratios of partition functions (that are also directly related to chemical equilibrium) can be computed by means of different techniques (a compromise between accuracy and efficacy). Thus the quality of free energy difference calculations is reflected by the choice of:

• a reliable molecular model (i.e Hamiltonian) describing the system's thermodynamics employed in energy and force calculations; more CPU time-demanding, and exact, is quantum-mechanical description (such as Free Energy Perturbation i.e FEP, Thermodynamic Integration i.e TI, or the non-equilibrium statistical mechanics approach by Jarzynski[129]), whereas more approximate approaches employ empirical force fields (such as Linear Interaction Energy i.e LIE[130], or the Molecular Mechanics Poisson-Boltzman/Generalized-Born Surface Area i.e MM/PB(GB)SA approach).

• a sampling method that is employed to generate an ensemble of representative configurations (such as MD or Monte Carlo methods)

2.4.2 MM/PBSA and MM/GBSA method:

The key objective of the approximate binding free energy MM/PBSA method,[131] and its complimentary MM/GBSA method,[132] is to calculate the free energy difference between two states which generally represent the bound and unbound state of two solvated molecules, or eventually to compare free energy of two different solvated conformations of the same molecule. Representative snapshots from an ensemble of conformations obtained via MD simulation are used for the calculation (stripped from the explicit solvent), yielding an average of the energies. The binding free energies of macromolecular association are computed by the means of a thermodynamic cycle (Figure 2.3), combin-

ing molecular mechanics energies (so-called gas phase energy contributions that are independent of the chosen solvent model) with implicit solvent approaches.

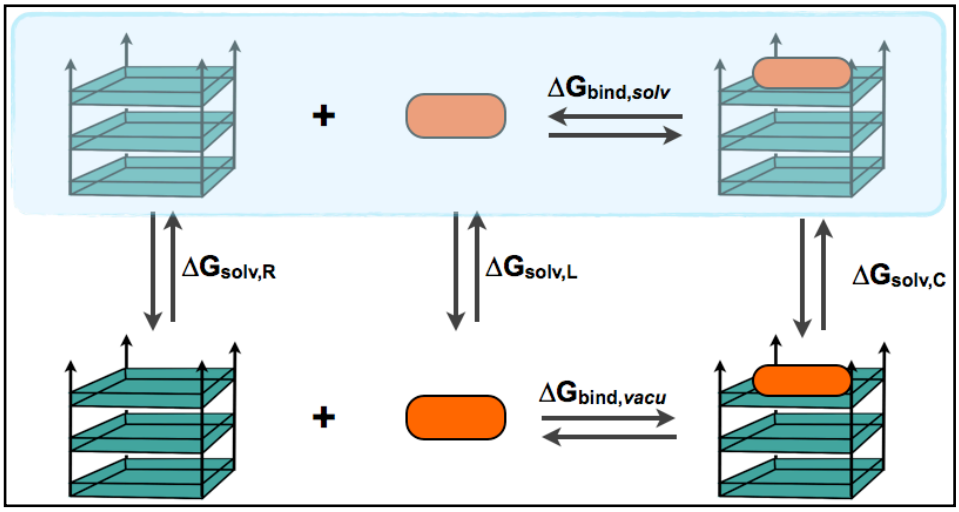


Figure 2.3: MM/PB(GB)SA thermodynamic cycle.
Molecular mechanics energies (gas phase energy contributions) that are independent of the chosen solvent model, are combined with implicit solvent approaches. The calculations are performed for all three systems: the receptor, ligand and the complex.

The free energy of binding is then calculated as given by equations 2.9, 2.10, which can be further broken down into individual energy terms (equation 2.11, 2.12, 2.13):

$$\Delta G_{bind} = \Delta H - T\Delta S \quad (2.9)$$

$$\Delta G_{bind} = (\Delta E_{MM} + \Delta G_{SOL}) - T\Delta S \quad (2.10)$$

$$\Delta E_{MM} = (E_{MM}^{complex} - E_{MM}^{receptor} - E_{MM}^{ligand}) \quad (2.11)$$

$$\Delta G_{SOL} = (\Delta G_{SOL}^{complex} - \Delta G_{SOL}^{receptor} - \Delta G_{SOL}^{ligand}) \quad (2.12)$$

$$\Delta S = (S^{complex} - S^{receptor} - S^{ligand}) \quad (2.13)$$

where ΔH is the enthalpic contribution to binding energy, ΔE_{MM} is the difference in average molecular mechanics energy, while the ΔG_{SOL} term accounts for the solvation free energy (comprising both polar and non-polar component); T is the temperature and ΔS is a change in entropy. Thus the binding free energy of two molecules can be declared as the sum of an intermolecular energy (evaluated using a MM force field), a solvation free energy energy term and an entropic term.

Solvation free energies are calculated by either solving the linear, more accurate, Poisson Boltzman (PB) equation, or more approximate, and computationally more effective Generalized Born (GB) equation. The calculations performed for each of the three states (complex, receptor and ligand) provide the electrostatic contributions to the solvation free energy, while the hydrophobic contributions are calculated by means of an empirical term (i.e surface tension energy is related to the solvent accessible surface area, SASA). Entropy contributions to the total free energy may be estimated either by quasi-harmonic analysis[133] or by using normal mode analysis. However, while the enthalpic contribution to binding free energy (i.e the change in enthalpy ΔH) is easily obtained from the molecular mechanics forcefield itself, together with the complex solvation effects that can be accounted for by an implicit solvation model, the entropic contribution to binding free energy (i.e the change in configurational entropy) has been more challenging to calculate.[133]

The Poisson-Boltzmann (PB) equation is the core of one of the most popular implicit solvent models, as continuum electrostatic approximations are based on numerical solutions to the non-linear PB equation (equation 2.14), which comprises the Gauss' law (or Poisson equation) for electrostatic potential with the Boltzmann charge density. A linearized approximation to the PB equation is employed within the MM/PBSA method to represent the ionic nature of a biomacromolecule (solute) immersed in aqueous solution with counter-ions.

$$\nabla \cdot [\varepsilon(x)\, \nabla\Phi(x)] = -4\pi\rho_{solute}(x) - 4\pi\sum_{n=i} q_i\, c_i \exp\left[-q_i\, \Phi(x)/k_B T\right] \qquad (2.14)$$

The Generalized Born (GB) model is then an approximation to the linearized PB equation. MM/PBSA (and MM/GBSA) methods have recently been widely employed in binding free energy calculations of various protein-ligand systems,[134] as well as probing protein interfaces for binding affinity.[135] They have been exploited in the ranking of ligand binding affinities, which is important in particular in the drug discovery field. However limitations of the method have been reported in terms of MM/PBSA performance for structurally-diverse structures[136] or poor approximation of the entropy loss which has been challenging to compute in a number of cases, or often simply ignored.

2.5 VIRTUAL SCREENING TECHNIQUES: MOLECULAR DOCKING

The expression virtual screening (VS), which emerged in the late 1990s as an alternative to experimental high-throughput screening (HTS), describes the use of computational algorithms and models for the identification of novel bioactive molecules. Over the years, it has become an important component of the *in silico* search for hit and lead compounds and their optimization. Numerous such methods and applications have been described and widely reviewed.[137,138] VS can be broadly divided into two main categories, namely structure-based VS (i.e molecular docking, structure-based pharmacophore modeling), which utilizes the 3D structure of a biological target, and ligand-based (similarity-based) VS, where the structure-activity data from a set of known active compounds are employed.[139]

Molecular docking is a term used for a large-scale computational plan that aims to find the 'best' matching between two molecules: a receptor and a ligand. Given the atomic coordinates of two molecules, docking predicts their "correct" bound association (binding pose),[140] and estimates the strength of the interaction. There are three crucial components in docking:

(1) representation of the system

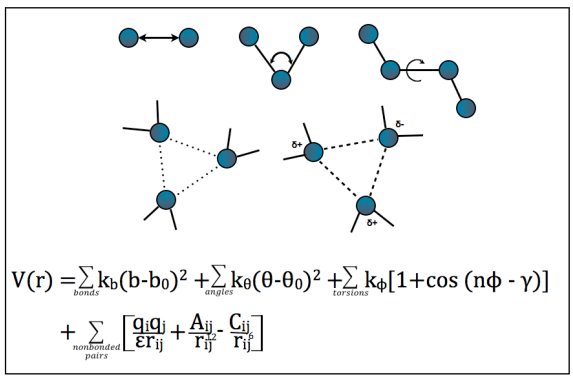
(2) conformational space search via a search algorithm

(3) ranking of potential solutions using the scoring function/energy function.

Molecular surfaces can be described by either a mathematical model (e.g. geometrical shape descriptors) or by a grid; but they can also be described by a molecular frame treatment - rigid or flexible. Incorporation of ligand and receptor flexibility to some extent, is currently required for accurate docking since the simplistic rigid 'lock-and-key' model of ligand receptor interaction is not adequate.[141] Thus rotational, translational and conformational degrees of freedom of the ligand are sampled at each docking run, and in some cases, further conformational degrees of freedom of residues within the ligand binding site are taken into account. The search algorithm generates a set of candidate geometries for a particular ligand, and then the configurations are evaluated (ranked) against each other using scoring functions, proposing the top-scoring pose at the global

minimum. The complexes are eventually re-scored in the end of the search to rank the hits in terms of their binding energy (affinity) for the receptor of interest. Despite significant progress that has been made in the generation of potential ligand poses by automated molecular docking, and the experimental poses can be predicted with reasonable accuracy,[142] there are still a number of pitfalls in the existing virtual screening methods (described in recent reviews by Schneider[143] or Scior[144] *et al*), that need to be addressed. These are related to the scoring functions used for binding affinity predictions for a diverse set of molecules, followed by ranking their binding potencies;[142] namely oversimplification of the energy terms usually employed in the scoring functions, such as entropy contributions,[145] specific non-covalent interactions that are not commonly included in the scoring functions,[146] water-mediated hydrogen bonds,[147] or the target flexibility upon ligand binding. Thus more alternative approaches to docking are desirable, and will be discussed as a part of the work described here (CHAPTER 5 & 6).

‘INTERMEZZO’



$$V(r) = \sum_{bonds} k_b(b-b_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 + \sum_{torsions} k_\phi[1+\cos(n\phi - \gamma)] + \sum_{\substack{nonbonded \\ pairs}} \left[\frac{q_i q_j}{\varepsilon r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}\right]$$

“I don’t demand that a theory correspond to reality because I don’t know what it is. Reality is not a quality you can test with litmus paper. All I’m concerned with is that the theory should predict the results of experiment.”
(Stephen Hawking)

'Overture'

A key aspect of many biological processes involving nucleic acids and their complexes, directly relates to their sequence-specific structure and motion. This motion (sequence-dependent flexibility and dynamics), brings an additional level of complexity beyond structure, that is essential for interpreting the function of nucleic acids (and their complexes). Molecular dynamics (MD) simulations with empirically-derived force fields is a technique, that can in principal, provide a complete theoretical description of a structure in motion. A critical aspect of such MD simulation is the choice of potential energy function (i.e its functional form, force field), that directly affects the modeling efficacy, as the the accuracy of modeling depends on the correctness of the potential.[148] Nucleic acids are very complex and flexible systems (with particular charge distribution along the backbone). DNA backbone conformation is defined by six different torsion angles, and as a result, the conformational space of nucleic acids is significantly more complex than the conformational space of proteins, hence requiring longer simulation times for the system relaxation.[149] Further consideration for simulations of DNA complexes, lies in the force field, which needs to be balanced in terms of pair interactions involving all combinations of solvent and the solute atoms.[150]

Over the past decade or so, there has been a massive improvement in the empirically-derived molecular mechanics force fields, tremendous gain in computational power, and hence in parallelization of the the simulation programs. This in turn, has contributed to great increase in reliability of MD simulations applied to nucleic acids and their complexes, and to the ability of not only reproducing the experimentally-obtained data, but also providing an insight beyond the experiment (i.e duplex,[151] triplex,[152] quadruplex[83,84] or supercoiled DNA[153] simulations). However, as outlined above, the reliability of simulations of DNA complexes, is directly related to the careful choice of the most advanced/improved force field parameters (and their combinations) that are currently available, together with an effective MD code, that will allow optimal choice of simulated conditions and reasonable cost of computational time.

==========

CHAPTER 3:
Porting AMBER force field parameters for nucleic acids, *parmbsc0*, into GROMACS and their validation, and introducing parameters for phosphorylated tyrosine residue

==========

3.1 BACKGROUND:

There is a good diversity in the software packages that may be used today for MD simulations of biomolecular systems. Each widely used and reviewed MD code features certain specific advantages over the others, which need to be considered prior to starting the MD simulation. However, the majority of these codes can employ force field, structure, and trajectory file formats that were originally introduced and/or produced in other programs. The choice of a force field to be used for the MD simulation should also be thought of, reflecting the nature of the simulated system.

3.1.1 Using GROMACS to simulate DNA complexes

GROMACS[154,155] an acronym for GROningen MAchine for Chemical Simulation, is presented as a versatile suite for molecular dynamics simulations of biomolecules, systems with complicated bonded interactions. GROMACS does not have a force field of its own, but it supports force fields such as GROMOS96,[156] OPLS-AA,[157] or AMBER.[158] Like AMBER, GROMACS has evolved from the initial version of CHARMM,[159] but it has since been fully developed and distributed as a free software, under the terms of the GNU (General Public License, http://www.gnu.org/). A great emphasis on algorithmic optimization has been introduced into GROMACS, resulting in its exceptional performance with respect to the other MD programs; with the ability to run efficiently on desktop computers, and in parallel employing standard MPI communication (both CPUs and GPU-accelerated MD simulations are supported). All

GROMACS file formats are plain-text based, so they are human-readable and editable in a situation when new residue/parameters need to be added to an existing force field. There are three different types of topology files utilized by GROMACS:

(I) *top* system topology; defines the entire system topology, either directly or by including .itp files.

(II) *itp* include topology; defines individual, or multiple, components of a topology as a separate file, and it is force field-specific

(III) *rtp* residue topology; the file contains the default interaction type for the bonded interactions and residue entries (i.e atoms).

Such files are needed to define a GROMACS topology for a macromolecule contained in a PDB file. Furthermore, each force field has a defined set of atom types with characteristic static properties, such as name/number, and mass (atomtypes.atp file). The charge is then listed in the residue topology file.

AMBER stands for Assisted Model Building with Energy Refinement, and it refers to two assets: a suite of separate programs that are designed to work together throughout the three main steps of a system preparation, molecular dynamics simulation, and trajectory analysis; and a set of molecular mechanical force fields for the simulation of biomolecules.[160] Being continuously improved/developed since the 1980s, AMBER (together with CHARMM) is one of the more established and respected MD packages.[118] PMEMD code, a recent addition to AMBER, provides scaling on profoundly parallel platforms (up to 64 cores), as a response to the massively increasing computer resources currently available.

3.1.2 Force field choice for MD simulations of DNA complexes

The force fields implemented in AMBER and CHARMM are the most widely used and reviewed force fields. In particular, AMBER force fields have proven to be the best choice for nucleic acid (both canonical and noncanonical) MD simulations in terms of accuracy of reproduction of their structural and dynamic properties. It also successfully reproduced hydrogen bond and stacking interactions which were hypothesized by

high-level QM data. Moreover, a number of systematic studies suggest that the AMBER force field is physically meaningful and retains a proper balance between intramolecular and intermolecular forces, and they also demonstrated good performance in extreme environments.[161] However, in MD simulations of DNA extended beyond the time-scale of 10-ns, tremendous α/γ transitions to the *gauche*$^+$, *trans* geometry, which subsequently introduced serious distortions in DNA at 50-ns trajectories, were reported.[162] This disturbing effect, was later confirmed by other research groups, as a general sequence-independent issue of parm94[163] and parm99[164] simulations. Since the α/γ torsional is important for the description of the low-twist conformations, which in turn may be critical during the protein-DNA recognition processes, there was a need for the correction (with perspective to the longer explicit MD simulation time that are currently possible).

Full reparametrization of the α/γ torsional term in nucleic acids was carried out by Perez and co-workers, leading to a new AMBER force field for nucleic acids, known as *parmbsc0*.[161] Based on AMBER-parm99,[164] this force field improves the representation of the α/γ conformational space of nucleic acids. In principle, the standard nomenclature employing the O3'-P-O5'-C5' (i.e. α torsion) and O5'-C5'-C4'-C3' (i.e γ torsion) to represent the α and γ dihedrals is retained as in the original parm99 force field (Figure 3.1), and a new atom type (CI) was introduced, and assigned to C5' carbon (in order to prevent alteration of the other conformational profiles).

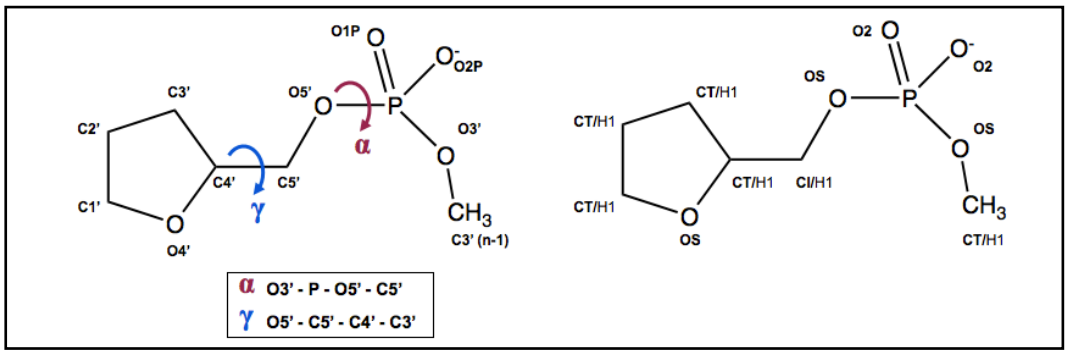


Figure 3.1: α and γ dihedral term representation as defined in AMBER *parmbsc0.*[161]
Schematic representation of the molecular model of the sugar-phosphate component of the nucleotide, that was used to define α and γ torsions and port them into to GROMACS. The atom-type definition is also displayed.

3.2 AIMS

In its simple form, a biomolecular force field is composed of two distinct constituents, which describe the interaction between particles (i.e atoms);

(1) the set of equations (i.e the potential functions) employed to generate the potential energies and their derivatives, the forces

(2) the parameters, that are used in this set of equations

Exceptions can be found, but it is typical for a force field to be purely additive, and transferable between the simulation programs, provided that the code may be slightly modified due to the discrete differences in torsions, units etc.

Here, the ultimate goal is to

• successfully port AMBER *parmbsc0*[161] parameters for nucleic acids into GROMACS v 4.5, for the subsequent molecular dynamics simulations of protein-DNA complexes and G-quadruplex nucleic acids complexes with small-molecule ligands, that are the scope of the following chapters described in 'PART 1' (protein-DNA MD simulations) and 'PART 2' (G-quadruplex/ligand MD simulations) respectively.

• additionally introduce AMBER parameters for phosphorylated tyrosine residue, reported by Homeyer[165] *et al*, into GROMACS v 4.5,[154,155] which was employed for MD simulations of phosphorylated STAT3-DNA complex, discussed within the chapters in 'PART 1'.

It should be noted, that while the MD simulations described in this thesis, were carried out between 2010 and 2012 (July), ACPYPE,[169] a tool based on ANTECHAMBER[170] for generating automatic topologies and parameters in different formats was released and published. The ACPYPE tool was then employed here, in order to verify the reproducibility of the *parmbsc0* parameters that were previously manually ported into GROMACS.

3.3 CONVERSION AND VERIFICATION OF *parmbsc0* IN GROMACS

A detailed description of AMBER *parmbsc0*[161] parameters implementation into GROMACS, and its validation is demonstrated here. Reparametrized nucleic acid parameters of the AMBER parm99[164] force field, *parmbsc0*[161] parameters, were obtained from the frcmod file (parmbsc0.frcmod), which comprises all the specifications regarding the atom types, atom mass, and bonded (bonds, angles, dihedrals, improper dihedrals) and non-bonded terms. These parameters were then manually converted to GROMACS-accommodated AMBER-port parm99SB-ILDN, an improved protein side-chain torsion potential force field developed at Shaw's[166] research group (based on parm99SB,[167] a modification of parm99), by replacing the distinct nucleic acids parameters with *parmbsc0* parameters. The principles by which GROMACS accommodates the potential energy functions used in force fields were carefully followed, and subsequent validation of successful parameter conversion was carried out.

3.3.1 Introducing new atom types and parameters into an existing force field

Introduction of a new residue to an existing forcefield into GROMACS, or modification of an existing forcefield, such as this case, requires an alteration of several topology and parameter files, with respect to the following steps:

• A novel residue, or a modification to an existing residue, is added to the *.rtp file (i.e. dna.rtp) of the chosen force field, and to the residuetypes.dat file with the appropriate specification.

• If hydrogens are added with the new residue, a relevant entry should be made within the *.hdb file.

• New atom types need to be added into the atomtypes.atp and ffnonbonded.itp files

• If new bonded types (bonds, angles and/or torsions) are required, they need to be added into the ffbonded.itp.

• specbond.dat file needs to be updated, if the added residue involves special connectivity to other residues.

Prior to any modifications, the parm99SB-ILD AMBER-port forcefield directory (with all its parameter and topology files) implemented within GROMACS v 4.5.3[154,155] was copied from its original location to the working directory (preventing any accidental modifications to the original AMBER port) and saved using an alternate name.

The equation used for the potential energy, known as a functional form of a force field is outlined here in its simplified form (equation 3.1), as a summation of the bonded (bonds, angles, torsions) and non-bonded (Lennard-Jones and Coulombic) terms (equations 3.2 and 3.3):

$$V(r) = E_{bonded} + E_{non\text{-}bonded} \quad (3.1)$$

$$E_{bonded} = E_{bonds} + E_{angles} + E_{torsions} \quad (3.2)$$

$$E_{non\text{-}bonded\ pairs} = E_{VdW} + E_{electrostatic} \quad (3.3)$$

The following sections discuss the conversion of: (i) van der Waals non-bonded parameters; (ii) the bonded parameters conversion, with emphasis on the 4-body interactions; and (iii) subsequent verification of the successful conversion of *parmbsc0*[161] parameters.

3.3.2 Conversion of non-bonded parameters

In order for the new *parmbsc0* parameters to be recognized in GROMACS when topology files for the simulated DNA molecules are generated; a new entry for the CI atom type (assigned to C5'), together with its atom mass specification, was introduced into the atomtypes.itp file. Also the CT atom types assigned to C5' in the dna.rtp file were changed to CI. Subsequently the CI atom type parameter, and its atomic mass of 12.01, was introduced into the ffnonbonded.itp file, together with two required and atom type-specific van der Waals parameters; sigma (σ; i.e. van der Waals atomic radius) and epsilon (ε; i.e. the depth of potential energy well).

In the original AMBER *parmbsc0*[161] frcmod file the CI atom is given the following radius and well depth parameters, where the first number is the atomic radius in Angstrom, and the second term is the well depth in kcal/mol.

CI 1.9080 0.1094

Since the CI atom type retained the Van der Waals parameters as defined in parm99, where any sp^3 aliphatic carbon, such as C5' (Figure 3.1), is assigned as CT atom type, a line with CT atom type specifications was duplicated in GROMACS ffnonbonded.itp file, and the duplicate was renamed as CI, with the following parameters:

```
[ atomtypes ]
; name   at.num  mass  charge  ptype  sigma          epsilon
CI       6       12.01 0.0000  A      3.39967e-01  4.57730e-01
```

where 'ptype' is the particle type (atom), sigma is the radius in nm, and epsilon is in kJ/mol. Sigma and epsilon values are directly used in this instance of using AMBER force field in GROMACS; they (epsilon and sigma values) are combined to provide parameters for all pairs of atom types in the system, and subsequently $C_{ij}^{(12)}$ (i.e. constant A) and $C_{ij}^{(6)}$ (i.e constant C) atom pair-dependent parameters, which are employed in GROMACS internal Lennard-Jones potential equation for two atoms are calculated and used in the program's internal version of the LJ potential equation (details given in CHAPTER 2).

3.3.3 Conversion of bonded parameters - bonds and angles

Bonded interactions are based on a fixed list of atoms; those are not exclusively pair interactions, such as bonds (1-2 interactions), but they comprise angles (1-3 interactions), and torsions (1-4 interactions) as well, and they are represented/modeled via combination of potentials with specific parameters. Description of the overall "Class I" potential energy function (i.e force field), and its energy components and their parameters, is given in detail in CHAPTER 2. However in practice there is a close relation be-

tween the force fields and the simulation programs (codes) that implement them, in terms of distinct differences of force constant expressions, units, or the functional forms of the individual equations.[168] Hence the mathematical shapes of the individual AMBER force field bonded energy terms are outlined here, both in their 'native' AMBER, and 'adapted' GROMACS format, to demonstrate the discrete features and differences that need to be considered for successful *parmbsc0*[161] parameters conversion.

The harmonic potential representing bond length stretching between atomic pairs, where atoms are separated by one covalent bond, is outlined in its functional shape as implemented in AMBER and GROMACS in Table 3.1; while the harmonic potential term describing the alteration of bond angles from their ideal values is outlined in Table 3.2. The forms of the 2-body and 3-body interaction terms are corresponding between each other, and also between AMBER and GROMACS. However, there is a difference in the default units used by the two codes, and also AMBER does not use the factor ½ in harmonic potentials, which was accounted for in the *parmbsc0*[161] parameters conversion to GROMACS.

Table 3.1: Bond (1-2 interactions) potential function term, as implemented in AMBER and GROMACS.

FF	Energy term - BONDS	b_0 units	K_b units
AMBER	$E_{bond} = \sum K_b (b - b_0)^2$	Å	$kcal.mol^{-1}.Å^{-2}$
GROMACS	$V_b(b_{ij}) = ½ K_b (b - b_0)^2$	nm	$kJ.mol^{-1}.nm^{-2}$

Table 3.2: Angle (1-3 interaction) potential function term, as implemented in AMBER and GROMACS.

FF	Energy term - ANGLES	θ_0 units	K_θ units
AMBER	$E_{angle} = \sum K_\theta (\theta - \theta_0)^2$	°	$kcal.mol^{-1}.rad^{-2}$
GROMACS	$V_a(\theta_{ijk}) = ½ K_\theta (\theta - \theta_0)^2$	°	$kJ.mol^{-1}.rad^{-2}$

AMBER *parmbsc0*[161] ideal (i.e reference) bond length (b_0) and ideal bond angle (th_0) values, and their respective force constants K_b and K_θ are listed in Table 3.3 and 3.4, together with their corresponding values for the specific bonds/angles upon parameter conversion into GROMACS.

Table 3.3: Force field parameters describing the bonds in *parmbsc0*[161] force field.

BOND	AMBER		GROMACS	
	K_b	b_0	K_b	r_0
	[kcal.mol^{-1}.Å^{-2}]	[Å]	[kJ.mol^{-1}.nm^{-2}]	[nm]
CI-H1	340.0	1.090	284512.0	0.10900
CI-CT	310.0	1.526	259408.0	0.15260
OS-CI	320.0	1.410	267776.0	0.14100
OH-CI	320.0	1.410	267776.0	0.14100

The ideal bond length values (b_0) were multiplied by 10^{-1} to convert them from default Ångstrom to nm units (which are used by GROMACS); while the bond force constants (K_b) were multiplied by (i) 4.184, to convert kcal. mol^{-1} to kJ. mol^{-1}, then by (ii) 100 to convert Å^{-2} to nm^{-2} and finally by (iii) 2 to account for the force constant factor used by GROMACS.

Table 3.4: Force field parameters describing the angles in *parmbsc0*[161] force field.

ANGLE	AMBER		GROMACS	
	K_θ	th_0	K_θ	th_0
	[kcal.mol^{-1}.rad^{-2}]	[°]	[kJ.mol^{-1}.rad^{-2}]	[°]
H1-CI-CT	50.0	109.50	418.400	109.50
H1-CI-H1	35.0	109.50	292.880	109.50
CI-CT-H1	50.0	109.50	418.400	109.50
CI-CT-OS	50.0	109.50	418.400	109.50
CI-CT-CT	40.0	109.50	334.720	109.50
OS-CI-H1	50.0	109.50	418.400	109.50
OS-CI-CT	50.0	109.50	418.400	109.50
P-OS-CI	100.0	120.50	836.800	120.50
OH-CI-H1	50.0	109.50	418.400	109.50
OH-CI-CT	50.0	109.50	418.400	109.50
HO-OH-CI	55.0	108.50	460.240	108.50

The AMBER values for ideal (i.e equilibrium) bond angle (th_0) were kept identical since both codes use degrees for angle units, and subsequently the angle force constants were multiplied by 4.184 (kcal. mol^{-1} to kJ. mol^{-1}), and by a factor of 2 to obtain corresponding angle force constants in GROMACS format.

3.3.4 Conversion of bonded parameters - dihedrals

The term describing dihedral (torsion) angle twisting is represented by a cosine function (periodic potential), to model the steric barriers (i.e rotation barrier height) between atoms I-J-K-L, separated by three covalent bonds. The related rotational motion is then described by a dihedral angle, hence the dihedral parameters, together with the atomic charges van der Waals parameters, are the primary determinants of the relative conformational energies of a molecule.

In AMBER, torsion potential formula parameters PK (one half of the rotation barrier magnitude), IDIVF (total number of torsions about a single bond that the potential applies to), PN (periodicity of a particular topology about the dihedral of interest, i.e the number of potential barriers) and PHASE (phase shift), are used to define the dihedral potential energy function. They are typical for each bonded series of atoms I-J-K-L. The torsional energy terms as implemented in AMBER and GROMACS codes are outlined in Table 3.5.

Table 3.5: Torsion angle potential function term, as implemented in AMBER and GROMACS.

FF	Energy term - TORSIONS	units
AMBER	$E_{dihedral} = \sum V_n/2\ [1 + \cos(n\phi - \gamma)]$	Vn/2 [kcal.mol^{-1}]
formula	$E_{tors} = (PK/IDIVF) * [1 + \cos(PN\phi - PHASE)]$	
GROMACS	$V_d(\phi_{ijkl}) = K_\phi\ (1 + \cos(\ n\phi - \phi_s)$	Kϕ [kJ.mol^{-1}]

To illustrate the nature of dihedral parameters in AMBER, and their subsequent conversion to GROMACS, we will use the first dihedral in Table 3.6 as an example. In case of the X-CI-OS-X dihedral, two "wild-cards" X are used for atoms attached to the central C-O atoms connected by a single bond. The IDIVF=3 suggests how many times the dihedral would be defined, if all the atoms had been expressed explicitly, without the wildcard X (if all dihedral atoms were explicitly expressed), IDIVF=1. PK is equal to one-half of the barrier (magnitude) height, which corresponds to the term Vn/2 that is used in the primary AMBER dihedral potential term (Table 3.5), and is subsequently divided by IDIVF factor, to represent a torsional barrier for a specific dihedral case. In GROMACS, this is represented by force constant Kϕ. Periodicity (multiplicity) n=3

defines the number of potential barriers as the C-O bond (CI-OS) is rotated from -180 to 180 degrees, and PHASE=0 suggest that there is an energy maximum at 0 degrees (while for energy minimum at 0 degrees PHASE=180), and that the potential energy barriers can be reproduced by the truncated Fourier series with no phase shift needed.

Based on the principles outlined above, the AMBER dihedral parameters conversion to GROMACS was done by dividing the PK (i.e Vn/2) value by factor IDIVF, and subsequent multiplication by 4.184 (kcal. mol^{-1} to kJ. mol^{-1}) to obtain Kϕ value; the phase angles (γ, ϕ_0 respectively) and absolute periodicity values (n) remained the same.

Table 3.6: Force field parameters describing the α/γ torsions in *parmbsc0*[161] force field.
The X-CI-CT-X AMBER dihedral (highlighted in *cyan*) was replaced by four explicit dihedrals (highlighted by *dark grey*) to be used in GROMACS.

DIHEDRAL	AMBER					GROMACS			
	tor # (IDIVF)	Vn/2 (~PK) [kcal.mol^{-1}]	γ (~PHASE) [°]	n (PN)	torsion	fn	ϕ_s [°]	Kϕ [kJ.mol^{-1}]	n
X-CI-OS-X	3	1.150	0.0	3.0		9	0.0	1.60387	3
X-CI-OH-X	3	0.500	0.0	3.0		9	0.0	0.69733	3
X-CI-CT-X	9	1.400	0.0	3.0		9	0.0	0.65084	3
OH-CI-CT-OS						9	0.0	0.65084	3
OS-CI-CT-OS						9	0.0	0.65084	3
H1-CI-CT-H1						9	0.0	0.65084	3
H1-CI-CT-CT						9	0.0	0.65084	3
CT-OS-CT-CI	1	0.383	0.0	-3.0		9	0.0	1.60247	3
CT-OS-CT-CI	1	0.100	180.0	2.0		9	180.0	0.41840	2
H1-CI-CT-OS	1	0.250	0.0	1.0		9	0.0	1.04600	1
H1-CI-CT-OH	1	0.250	0.0	1.0		9	0.0	1.04600	1
H1-CT-CI-OS	1	0.250	0.0	1.0		9	0.0	1.04600	1
H1-CT-CI-OH	1	0.250	0.0	1.0		9	0.0	1.04600	1
CI-CT-CT-CT	1	0.180	0.0	-3.0		9	0.0	0.75312	3
CI-CT-CT-CT	1	0.250	180.0	-2.0		9	180.0	1.04600	2
CI-CT-CT-CT	1	0.200	180.0	1.0		9	180.0	0.83680	1
OS-P -OS-CI	**1**	**0.185181**	**31.79508**	**-1.0**	**alpha**	9	**31.79508**	**0.774797**	**1**
OS-P -OS-CI	**1**	**1.256531**	**351.95960**	**-2.0**	**alpha**	9	**351.95960**	**5.257326**	**2**
OS-P -OS-CI	**1**	**0.354858**	**357.24748**	**3.0**	**alpha**	9	**357.24748**	**1.484726**	**3**
OH-P -OS-CI	**1**	**0.185181**	**31.79508**	**-1.0**	**alpha**	9	**31.79508**	**0.774797**	**1**
OH-P -OS-CI	**1**	**1.256531**	**351.95960**	**-2.0**	**alpha**	9	**351.95960**	**5.257326**	**2**
OH-P -OS-CI	**1**	**0.354858**	**357.24748**	**3.0**	**alpha**	9	**357.24748**	**1.484726**	**3**
CT-CT-CI-OS	**1**	**1.178040**	**190.97653**	**-1.0**	**gamma**	9	**190.97653**	**4.928919**	**1**
CT-CT-CI-OS	**1**	**0.092102**	**295.63279**	**-2.0**	**gamma**	9	**295.63279**	**0.385355**	**2**
CT-CT-CI-OS	**1**	**0.962830**	**348.09535**	**3.0**	**gamma**	9	**348.09535**	**4.028481**	**3**
CT-CT-CI-OH	**1**	**1.178040**	**190.97653**	**-1.0**	**gamma**	9	**190.97653**	**4.928919**	**1**
CT-CT-CI-OH	**1**	**0.092102**	**295.63279**	**-2.0**	**gamma**	9	**295.63279**	**0.385355**	**2**
CT-CT-CI-OH	**1**	**0.962830**	**348.09535**	**3.0**	**gamma**	9	**348.09535**	**4.028481**	**3**

The one exception of the converted AMBER dihedrals was a dihedral defined around the CI-CT bond using two "wild cards" X (highlighted in *cyan*, Table 3.6). Upon numerous unsuccessful attempts to reproduce the dihedral potential energy employing the X-CI-CT-X dihedral, this dihedral was replaced by four explicitly defined dihedrals[1]. Namely: (1) OH-CI-CT-OS, (2) OS-CI-CT-OS, (3) H1-CI-CT-H1, and (4) H1-CI-CT-CT, with the phase angle and barrier height values corresponding to the original X-CI-CT-X dihedral. This was done to prevent a misinterpretation of other explicitly defined dihedrals with corresponding CI-CT single bond at positions J-K that were further defined by AMBER (such as H1-CI-CT-H1 and H1-CI-CT-CT), and their confusion with X-CI-CT-X dihedral options. All *parmbsc0*[161] dihedrals in GROMACS were assigned as type 9 function types, which allow multiple potential functions to be applied automatically to a single dihedral (as defined in the system topology file for bonded interactions ffbonded.itp).

3.3.5 Testing and validation of the *parmbsc0* force field ported into GROMACS

The AMBER *parmbsc0*[161] parameters conversion to GROMACS was tested and verified by running a single point energy (SPE) molecular dynamics simulations of 9-bp DNA helix, both in AMBER and GROMACS, employing the new improved *parmbsc0* force field (i.e the *parmbsc0*[161] parameters, frcmod and lib files, loaded into parm99SB), and the standard AMBER parm99SB[167] force field (used as a control). A flowchart demonstrating the testing of *parmbsc0*[161] force field ported into GROMACS, carried out to validate the ability of the newly created parmbsc0-port to reproduce the "original" AMBER test simulation results, is graphically outlined in Figure 3.2. Three sets of corresponding test SPE simulations were performed as follows:

• in SANDER, (AMBER main program for MD simulations)

• in GROMACS, with AMBER topologies 'translated' into GROMACS topologies by ACPYPE,[169] a tool based on ANTECHAMBER[170] for generating automatic topologies and parameters in different formats

• in GROMACS, employing the AMBER-ports

[1] I acknowledge a helpful discussion with Dr. Ondrej Kroutil regarding the X-CI-CT-X dihedral.

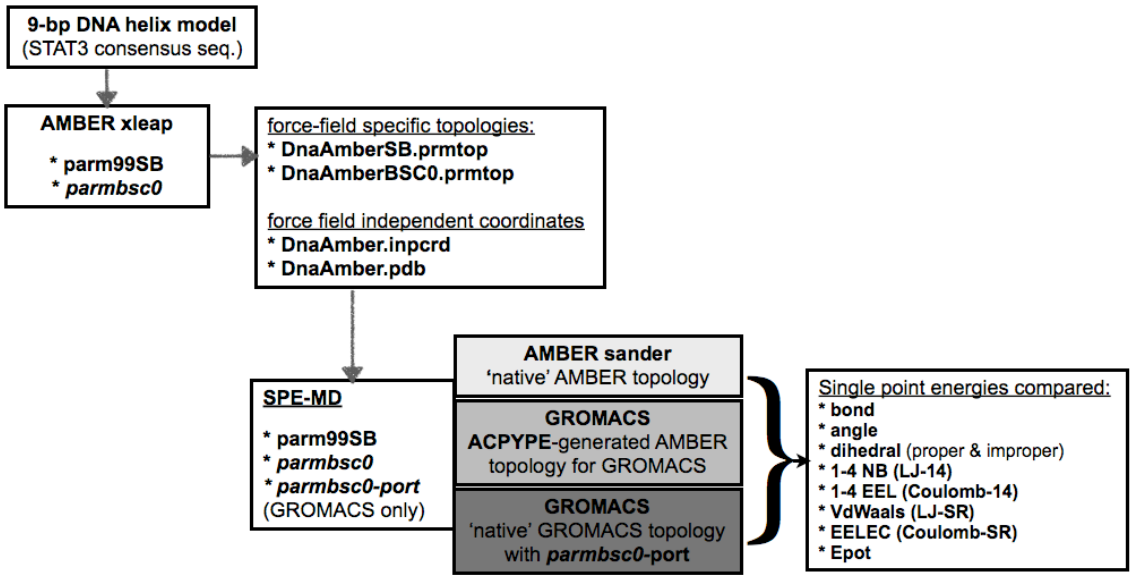


Figure 3.2: Flowchart demonstrating the testing and validation of the *parmbsc0*-port for GROMACS. 9-bp DNA helix topologies and coordinates were obtained through xleap, to be used for SPE test simulations in AMBER, ACPYPE[169] via GROMACS, and GROMACS itself, employing parm99SB,[167] *parmbsc0*[161] and parmbsc0-port manually created for GROMACS.

A model of *ds*DNA helix was obtained from the STAT3-DNA complex, by extracting the 9-bp STAT3 consensus sequence,[37] and saving it as a new PDB entry. This "test structure" was then processed through the xleap utility in AMBER, to obtain the topology and coordinate files, which were generated by employing the parm99SB[167] force field, and subsequently *parmbsc0*[161] force field, that was previously loaded into xleap in separate 'frcmod' and 'lib' files. Two sets of force field-specific topologies were obtained and saved in AMBER prmtop files; also coordinates (which are force field independent) in AMBER inpcrd file, and standard PDB file were output. The individual x-y-z coordinate entries within the AMBER coordinates file (inpcrd) were rounded to three decimal points, to match the coordinate entries of the PDB file, and ensure the consistency of the results, since these particular coordinate files were used for all the subsequent test simulations (including those in GROMACS). Corresponding input files for the SPE MD simulations were generated both for AMBER (sander) and GROMACS, with no constraints, no cutoff values for non-bonded interactions, and no solvent models applied. Three sets of SPE test simulations were then carried out according to the following scenarios, with their results summarized in Table 3.7.

Table 3.7: Testing and verification of AMBER *parmbsc0* parameters converted to GROMACS.
Single point energy test simulations (without any solvent) of 9-bp DNA helix were performed with AMBER parm99SB and *parmbsc0* parameters in native AMBER; in GROMACS via ACPYPE employing AMBER-generated topologies; and in GROMACS using the the manually converted *parmbsc0* parameters, introduced into AMBER port in GROMACS v. 4.5

9bp *ds*DNA	AMBER 99SB	AMBER BSC0	AMBER verif.	ACPYPE 99SB	ACPYPE BSC0	ACPYPE BSC0	ACPYPE BSC0	pdb2gmx 99SB	pdb2gmx 99SB-BSC0
	[kcal/mol]		[kJ/mol]	dihe.**3**, imp.**1**	dihe.**3**, imp.**1**	dihe.**9**, imp.**4**	dihe.**1**, imp.**1**	dihe.**9**, imp.**4**	dihe.pr.**9**, imp.**4**
BOND	48.9178	48.9178	204.67	204.67	204.67	204.67	204.67	204.67	204.67
ANGLE	422.8055	422.8055	1769.02	1769.02	1769.02	1769.02	1769.02	1769.02	1769.02
DIHED	384.4704	395.3448	1608.62 // 1654.1	0.801782	129.668	1524.45	1654.12	1607.82	1653.32
Ryckaert-Bell	---	---	---	1607.82	1524.45	---	---	---	---
Improper	---	---	---	---	---	129.668	---	0.80178	0.801782
Suma DIHED	384.4704	395.3448	1608.62 // 1654.1	1608.62	1654.12	1654.12	1654.12	1608.62	1654.12
1-4 NB (LJ-14)	184.7365	184.7365	772.937	772.935	772.935	772.935	772.935	772.935	772.935
1-4 EEL (Coulomb-14)	-1578.9788	-1578.9788	-6606.4473	-6606.42	-6606.42	-6606.42	-6606.42	-6606.41	-6606.41
VdWaals (LJ-SR)	-248.4538	-248.4538	-1039.5307	-1039.53	-1039.53	-1039.53	-1039.53	-1039.53	-1039.53
EELEC (Coulomb-SR)	1597.3073	1597.3073	6683.1337	6683.45	6683.45	6683.45	6683.45	6683.37	6683.45
Epot	810.8049	821.6794	3392.41 // 3437.9	3392.74	3438.24	3438.24	3438.24	3392.67	3438.17

(I) Two SPE simulations were completed in sander (AMBER), using the parm99SB and *parmbsc0* -generated topologies. Individual components of their bonded and nonbonded energy terms, such as bonds, angles, dihedrals (both proper and improper), Lennard-Jones and Coulomb interactions (both long-range and short-range), as well as the overall potential energy were extracted from the resulting output files and compared.

(II) The AMBER parm99SB and *parmbsc0* topology and coordinate files were processed via the ACPYPE tool, which "translated" them into GROMACS topology file formats, for the subsequent four SPE test simulations carried out in GROMACS v 4.5. The numerous ACPYPE-employing simulations were used mainly for understanding the effects of using different function types for the dihedral parameters, as the AMBER improper dihedrals are treated as proper dihedrals in GROMACS. Specifically, the dihedral functions explored were: (1) for proper dihedrals, Ryckaert-Bellemans function,[171] a type 3 function (which is internally used by GROMACS to compute Fourier dihedrals) was applied, together with a standard periodic proper dihedral type 1[155] function used to compute the AMBER improper dihedrals contribution. In case of *parmbsc0*[161] parameters, the *parmbsc0* dihedral contribution were also computed by function type 1 here; (2) then variant of the standard periodic type dihedral type 1, type 9[155] function (which is suggested by the GROMACS manual as useful when multiple potential functions are applied to a single dihedral), was combined with an improper dihedral periodic type 4, which is identical to type 1, but also distinguishes improper from proper dihedrals in the parameter section and in the output; (3) at last, both proper and improper dihedrals were assigned the type 1 function for the calculation. As for the AMBER simulations, individual components of their bonded and nonbonded energy terms, and the overall potential energy were extracted from the resulting output files and compared.

(III) Two SPE test simulations of the 9-bp DNA helix were carried out in GROMACS v 4.5, with the topologies generated by *pdb2gmx* GROMACS tool, employing the AMBER-port force fields. First, the parm99SB AMBER port, which is by default a part of GROMACS v 4.5 suite, was used; secondly, the customized AMBER-port comprising the *parmbsc0* parameters manually introduced into parm99SB-port, previously copied into the working directory, was employed. As in the previous two cases, the

individual energy components were compared between the two GROMACS simulations, but also among each other within all of the eight test simulations.

All test simulation results may then suggest the following conclusions:

• *parmbsc0*[161] parameters do indeed improve the treatment of α/γ torsions in nucleic acids simulations. As expected, the values of the energy components were corresponding between parm99SB and *parmbsc0* with the exception of dihedrals, and hence the overall potential energy. Improved dihedral potential and the total potential energy values were observed for the *parmbsc0* force field compared to the standard parm99SB.

• From the four simulations based on the ACPYPE-translated topologies from AMBER to GROMACS, a number of observations were made regarding the dihedral function types 1, 3, 4 and 9, which are by default implemented in GROMACS; (1) in terms of the summarized dihedral potential energy, using the dihedral function type 3 (Ryckaert-Bellemans function[171]), together with the periodic dihedral type 1, led to corresponding results, as when function types 9 and 4 were employed. However, when types 3 and 1 were combined their dihedral potential energy outputs within the Ryckaert-Bellemans and proper dihedral sections were found to be in the same proportion as the outputs of types 9 and 4 within the proper and improper dihedrals section. Applying the type 1 function solely, to both proper and improper dihedrals, led to a single dihedral potential energy output within the proper dihedral section. In general, replacing dihedral function types 3 and 1 with 9 and 4, or replacing them by type 1 only, did not make a difference with respect to the summarized dihedral potential energy (comprising proper, improper and Ryckaert-Bellemans functions). Furthermore, the total potential energy remained unchanged when the corresponding force fields were compared. This suggests that there is a variability in the individual user's preferences for the output choice, without affecting the total dihedral and total potential energy.

• Excellent agreement was achieved between the outcomes of the SPE test simulations with respect to the 'native' AMBER, ACPYPE-translated topologies for GROMACS, and also GROMACS, using the parmbsc0-port with the manually introduced *parmbsc0* force field.

• In summary, the *parmbsc0* parameter conversion for the use in AMBER-port in GROMACS v. 4.5 was successful.

3.4 CONVERSION OF AMBER PARAMETERS FOR PHOSPHORYLATED TYROSINE RESIDUE INTO GROMACS

Structural properties of phosphorylated proteins are important with respect to molecular recognition and regulatory processes, since protein phosphorylation is an important process governing key steps in cellular regulatory networks. Protein phosphorylation is a very specific process, affecting only side chains of particular protein residues (such as serine, tyrosine or threonine), leading to a formation of phosphodiester linkage between hydroxyl group oxide of amino acid residues, and phosphate group.[172] A phosphate group bound to an amino acid side chain has either a single- or double-negative charge, with respect to the surrounding environment. Understanding of the role played by this group in terms of protein structure formation (or catalytic events) is directly linked to the knowledge of whether the phosphate group is mono- or di-anionic.[173]

For the *in silico* studies of STAT3βtc:homodimer complexed with DNA, that are described in detail in CHAPTERS 4 and 5, introduction of additional AMBER force field parameters for phosphorylated tyrosine residue was required. A consistent set of publicly available AMBER force field parameters designed for phospho-amino acids was developed by Homeyer[165] *et al*, from which the parameters for phospho-Tyr with unprotonated phosphate group were employed here (Figure 3.3). The choice of dianionic charge for phospho-Tyr in STAT3 was based both on literature survey[173,174] and on biological experiments via personal communication.

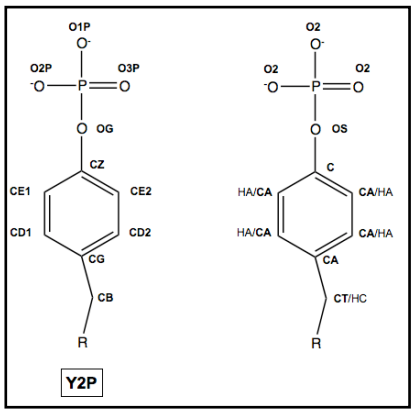


Figure 3.3: Schematic representation of the phospho-Tyr (Y2P) model with atom-types definition. This model of Y2P[165] was used to introduce new parameters for phosporylated Tyr into to GROMACS

3.4.1 Introducing phospho-Tyr residue into AMBER-port in GROMACS

The dianionic phospho-Tyr (Y2P) parameters[165] were obtained in the form of two AMBER force field parameters files; (1) frcmod file with the bonded parameters specifications, and (2) OFF library file, where atom names, atom types, and charges of the phospho-Tyr atoms were stored. The converted Y2P parameters were transferred into the AMBER-port of parm99SB-ILDN[166] in GROMACS (into a copy of the port saved in the working directory). Several distinct steps (principles) for converting/introducing new residue to AMBER-port in GROMACS v. 4.5, that are described above, were applied and followed here too. First, the new phosphotyrosine residue with dianionic charge, abbreviated as Y2P, was introduced into the residue topology file within the parm99SB-ILDN AMBER port. Individual atoms and their corresponding atom types were listed explicitly, together with their respective charges (Table 3.8) into the [atoms] section of the [Y2P] residue.

Table 3.8: Y2P residue atom name, atom type, and partial charges specification.
Atom names, and charges are corresponding between the original Y2P.frcmod file and the GROMACS residue topology file within the AMBER-port.

Y2P				
atom name	**atom type**	**charge [e]**	**atom #**	**atom type definition**
N	N	-0.516300	1	**N** -sp^2 nitrogen in amides
H	H	0.293600	2	**H** -H attached to N
CA	CT	0.141279	3	**CT** -any sp^3 carbon
HA	H1	0.027982	4	**H1** -H attached to aliphatic carbon with
CB	CT	-0.251171	5	one electron-withdrawing substituent
HB2	HC	0.081930	6	**HC** -H attached to aliphatic carbon with
HB3	HC	0.081930	7	no electron-withdrawing substituents
CG	CA	0.073446	8	**CA** -any aromatic sp^2 carbon
CD1	CA	-0.198084	9	
HD1	HA	0.112503	10	**HA** -H attached to aromatic carbon
CE1	CA	-0.254266	11	
HE1	HA	0.161606	12	
CZ	C	0.477009	13	**C** -any carbonyl sp^2 carbon
CE2	CA	-0.254266	14	
HE2	HA	0.161606	15	
CD2	CA	-0.198084	16	
HD2	HA	0.112503	17	
OG	OS	-0.557944	18	**OS** -sp^3 oxygen in ethers
P	P	1.380672	19	**P** -phosphorus in phosphates
O1P	O2	-0.943550	20	**O2** -sp^2 oxygen in anionic acids
O2P	O2	-0.943550	21	
O3P	O2	-0.943550	22	
C	C	0.536600	23	
O	O	-0.581900	24	**O** -sp^2 oxygen in amides

Bonds within the Y2P residue were further explicitly specified in the [bonds] section, followed by seven Y2P improper torsions in the [impropers] section. Angles and dihedrals were not explicitly listed here, as those are defined for all amino acids in the system topology file for bonded interactions (ffbonded.itp). Furthermore, all hydrogens within Y2P, together with their connectivity to other atoms, was specified in the hydrogen database file (aminoacids.hdb).

The newly-defined bonded parameters, one new bond, two angles, and one improper torsion angle (Table 3.9), were manually entered into the system topology file for bonded interactions, into the [bondtypes], [angletypes], and [dihedraltypes] sections.

Table 3.9: Bonded parameters for Y2P residue in AMBER, and upon conversion to GROMACS.

	AMBER			GROMACS			
BOND	**K_b**	**b_0**		**K_d**	**r_0**		
	[kcal.mol^{-1}.Å^{-2}]	**[Å]**		**[kJ.mol^{-1}.nm^{-2}]**	**[nm]**		
P-OS	525.0	1.610		439320.0	0.16100		
ANGLE	**K_θ**	**th_0**		**K_θ**	**th_0**		
	[kcal.mol^{-1}.rad^{-2}]	**[°]**		**[kJ.mol^{-1}.rad^{-2}]**	**[°]**		
C-OS-P	100.0	120.50		836.800	120.50		
CA-C-OS	70.0	120.00		585.760	120.00		
IMPROPER	**Vn/2** (~PK)	**γ** (~PHASE)	**n** (PN)	**fn**	Kϕ	ϕ_s	**n**
	[kcal.mol^{-1}]	**[°]**			**[kJ.mol^{-1}]**	**[°]**	
CA-CA-C-OS	1.100	180.0	2.0	4	4.60240	180.0	2

Since the ffbonded.itp topology file contains bonded interactions for both protein and nucleic acid systems, a P-OS bond for nucleic acids (r_0 = 0.16100 nm; K_d = 192464.0 kJ/mol) was already defined there. Despite the corresponding value of equilibrated bond length between the NA-specific and Y2P-specific P-OS bond, there is a real difference in their force constant values, where Y2P-specific K_d corresponds to the force constant value of P-O2 bond (K_d = 439320.0 kJ/mol). Hence a Y2P-specific flag for the P-OS bond (bond_Y2P_P_OS) was introduced at the relevant section of the residue topology file, and a corresponding bond flag was also used in the system topology file for bonded interactions (#define bond_Y2P_P_OS), together with the r_0 and K_d values. An explicitly-defined bond flag then ensured the use of correct parameters when the topology files of a simulated system were created. There were no identical angle entries for

the C-OS-P and CA-C-OS Y2P-specific angles in the ffbonded.itp topology file, so they were entered within the [angletypes] section. Lastly, the improper torsion specification with parameters were included into the [dihedraltypes] section, corresponding to the improper torsion listed for [Y2P] in the residue topology file (CA-CA-C-OS). The improper torsion was distinguished from proper dihedrals by assignment of the dihedral function 4, which was designed by GROMACS for this purpose.

The customized parm99SB-ILDN[166] AMBER-port with previously introduced *parmbsc0*[161] parameters for nucleic acids, and parameters for phosphorylated tyrosine residue[165] was then employed for all subsequent MD simulations of the STAT3-DNA complexes (CHAPTERS 4 and 5), and G-quadruplex DNA (CHAPTERS 6 and 7). The protein residue parameters did not interfere with the G-quadruplex DNA MD simulations, since only residues present in the PDB file of the molecular model are regarded for the system topology generation by GROMACS.

3.4.2 Retrospective testing and Y2P parameters validation in GROMACS v. 4.5

Following the force field parameters testing and validation procedure described above for *parmbsc0*, the phospho-Tyr parameters conversion from 'native' AMBER format (frcmod and lib files), and their subsequent administration into parm99SB-ILDN, and parm99SB AMBER-ports in GROMACS v 4.5, was retrospectively performed (with the latter being used as a control). A short tripeptide from the STAT3 sequence with the phosphorylated tyrosine residue (pTyr, pY), Pro-pTyr-Leu, and a control unphosphorylated sequence of Pro-Tyr-Leu, were used for the SPE test simulations.

Three sets of corresponding test SPE simulations were performed (Figure 3.4) :

• in SANDER, (AMBER main program for MD simulations) with Y2P parameters loaded into parm99SB and parm99SB-ILDN

• in GROMACS, with AMBER topologies 'translated' into GROMACS by ACPYPE[169]

• in GROMACS, employing the parm99SB[167] and parm99SB-ILDN[166] AMBER-ports, with manually introduced Y2P parameters

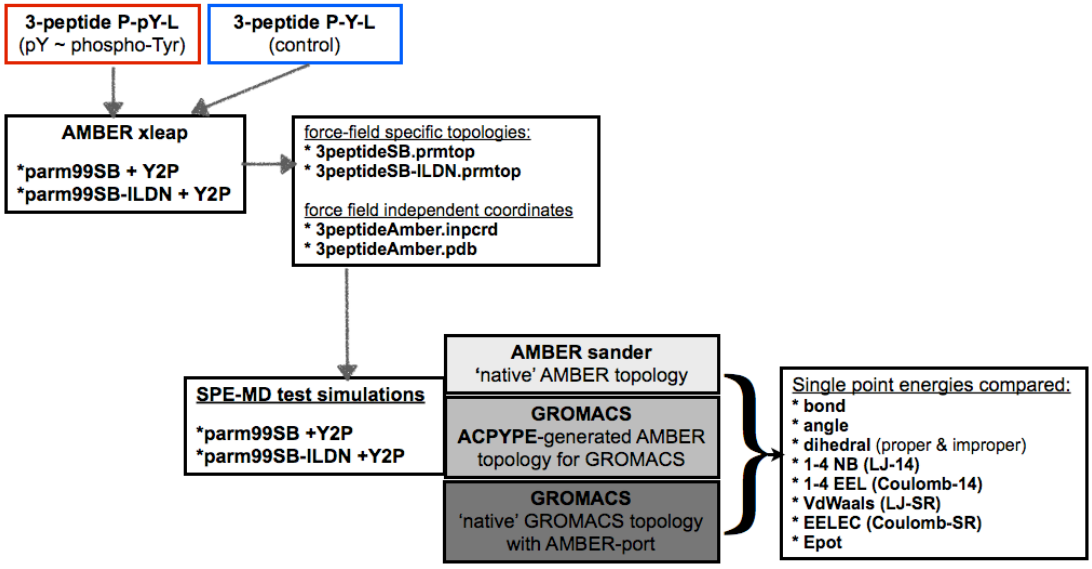


Figure 3.4: Flowchart demonstrating the testing and validation of Y2P parameters for AMBER-ports in GROMACS.
Tripeptide PpYL and PYL (control) topologies and coordinates were obtained through xleap, to be used for SPE test simulations in AMBER, ACPYPE via GROMACS, and GROMACS itself, employing parm99SB and parm99SB-ILDN force field with Y2P parameters. The individual potential energy components were then compared among each other.

The test SPE simulations, whose results are summarized in Table 3.10, were initially performed only for the Pro-pTyr-Leu tripeptide treated by parm99SB-ILDN force field with Y2P parameters, since that was the force field employed for the STAT3-DNA complex MD simulations. However, the comparison of the individual potential energy components displayed an inconsistency of ~ 3 kJ/mol in terms of the dihedral torsions, and the total potential energy, between the AMBER and GROMACS (i.e pdb2gmx) results (Table 3.10 a). This ~ 3 kJ/mol difference was not observed within the ACPYPE-employed simulations. To confirm or refute whether this inconsistency was due to faulty Y2P parameters conversion into GROMACS, a subsequent SPE test simulations were performed for the unphosphorylated Pro-Tyr-Leu tripeptide, also treated by the parm99SB-ILDN force field[166] (Table 3.10 b).

Table 3.10: Testing and validation of the Y2P parameters converted to AMBER-port in GROMACS
(a) Tripeptide PpYL and (b) PYL (used as a control) were employed in testing the Y2P parameters conversion and implementation within AMBER-port parmp99SB-ILDN in GROMACS v 4.5; AMBER-port parm99SB with implemented Y2P parameters was subsequently used as a 'control'.

(a) P-pY-L	AMBER 99SB	AMBER 99SBildn	AMBER verif.	ACPYPE 99SB	ACPYPE 99SBildn	pdb2gmx99SB	pdb2gmx99SBildn
	[kcal/mol]		[kJ/mol]	[kJ/mol] ; dihe.**9**, imp.**4**		[kJ/mol] ; dihe.**9**, imp.**4**	
BOND	15.0589	15.0589	63.0064	63.0038	63.0038	63.0037	63.0066
ANGLE	26.9868	26.9868	112.9128	112.91	112.91	112.91	112.912
DIHED	23.2224	23.2704	97.1625 // 97.3633	96.7775	96.9748	96.9775	93.9828
Improper	---	---	---	0.3845	0.3845	0.3847	0.3847
Suma DIHED	23.2224	23.2704	97.1625 // 97.3633	97.1620	97.3593	97.1622	94.3675
1-4 NB (LJ-14)	11.9363	11.9363	49.9415	49.9415	49.9415	49.9415	49.9415
1-4 EEL (Coulomb-14)	87.7384	87.7384	367.0975	367.097	367.097	367.096	367.096
VdWaals (LJ-SR)	-4.048	-4.048	-16.9368	-16.9371	-16.9371	-16.9371	-16.937
EELEC (Coulomb-SR)	-174.2107	-174.2107	-728.8976	-728.924	-728.924	-728.923	-728.923
Epot	-13.3159	-13.2678	-55.7137 // -55.5125	-55.7470	-55.5461	-55.7460	-58.5366

(b) P-Y-L	AMBER 99SB	AMBER 99SBildn	AMBER verif.	ACPYPE 99SB	ACPYPE 99SBildn	pdb2gmx99SB	pdb2gmx99SBildn
	[kcal/mol]		[kJ/mol]	[kJ/mol] ; dihe.**9**, imp.**4**		[kJ/mol] ; dihe.**9**, imp.**4**	
BOND	6.9247	6.9247	28.9729	28.9706	28.9706	28.9705	28.9705
ANGLE	14.5433	14.5433	60.8491	60.8491	60.8491	60.8491	60.8491
DIHED	22.9159	22.9639	96.0809	95.4996	95.7005	95.4996	92.7048
Improper	---	---	---	0.3802	0.3802	0.3802	0.3802
Suma DIHED	22.9159	22.9639	95.8801 // 96.0809	95.8798	96.0827	95.8798	93.0850
1-4 NB (LJ-14)	10.4703	10.4703	43.8077	43.8076	43.8076	43.8076	43.8076
1-4 EEL (Coulomb-14)	237.5450	237.5450	993.8882	993.8850	993.8850	993.8830	993.8830
VdWaals (LJ-SR)	-6.6941	-6.6941	-28.0081	-28.0081	-28.0081	-28.0081	-28.0081
EELEC (Coulomb-SR)	-250.7304	-250.7304	-1049.0559	-1049.09	-1049.09	-1049.09	-1049.09
Epot	34.9747	35.0227	146.3341 // 146.5349	146.2900	146.4910	146.2900	143.4950

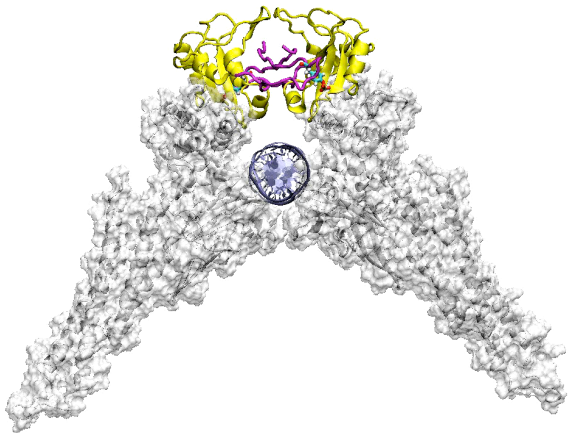
Even here, the corresponding ~3 kJ/mol inconsistency was observed, suggesting a GROMACS AMBER-port-specific error rather than faulty Y2P parameters conversion. A third round of corresponding SPE test simulations was carried out, with both tripeptides treated by parm99SB force field with Y2P parameters. The subsequent analysis revealed an agreement between AMBER, ACPYPE and GROMACS simulations, in terms of all explored potential energy components. Thus these three sets of SPE test simulations carried out in AMBER, ACPYPE and GROMACS revealed that the parm99SB-ILDN AMBER port implemented in GROMACS v 4.5 did not correctly reproduce AMBER dihedral parameters, leading to inconsistencies in the dihedral potential energy component and subsequently the total potential energy.

Further test SPE simulations were then performed, using a phosphorylated pentapeptide Pro-pTyr-Leu-Lys-Thr (also obtained from the STAT3 sequence). As with the tripeptides, a corresponding ~ 3 kJ/mol difference in the dihedral potential energy and total potential energy was observed between AMBER and GROMACS simulations, but not between AMBER and ACPYPE simulations (data not shown). This finding suggested that it might be tyrosine/phospho-tyrosine dihedrals not being correctly converted by AMBER-port parm99SB-ILDN in GROMACS v 4.5. To get a deeper insight into that problem, topologies for both the tyrosine and phospho-tyrosine residue (i.e bonds, angles, dihedrals and impropers) were extracted from their respective topology files generated by GROMACS and ACPYPE and compared. ACPYPE topology represents the AMBER topology as generated by sander, and it is in a human-readable format (unlike the original AMBER topology). All the bond, angle and improper torsion entries were corresponding between GROMACS and ACPYPE topology in terms of the Tyr and pTyr residue respectively. However, when comparing the individual dihedral torsions, one extra proper dihedral connecting atoms CZ-CE2-CD2-CG (i.e atom types C-CA-CA-CA; defined in GROMACS system topology file ffbonded.itp as X-CA-CA-X dihedral with K_d = 15.16700 kJ/mol) was found in the GROMACS-generated topology file for both phosphorylated and unphosphorylated tyrosine residues. This CZ-CE2-CD2-CG dihedral was not found in the ACPYPE topologies, so we may speculate, that this particular dihedral is causing the ~ 3kJ/mol discrepancy in parm99SB-ILDN AMBER-port compared to the parm99SB AMBER-port, and compared to the

original AMBER-generated topologies. Even though the parm99SB-ILDN[166] was assessed as the most suitable choice of force field for the protein part of the STAT3-DNA complex due to the improved parameters of amino acid side chains, namely Ile, Leu, Asp and Asn, there is a real requirement of improvement of its correct implementation within GROMACS v 4.5.

PART I

EXPLICIT SOLVENT MOLECULAR DYNAMICS STUDIES OF THE STAT3βtc-HOMODIMER:DNA COMPLEX



"Everything that living things do can be understood in terms of the jiggling and wiggling of atoms "
(Richard Feynman, 1963)

'Overture'

Signal Transducers and Activators of Transcription (STAT) proteins are a group of latent cytoplasmic transcription factors that relay signals from the plasma membrane (in response to stress, cytokines and growth factor signalling) to the nucleus.[30] The inappropriate activation of STAT is linked to various inflammatory diseases (i.e Hyper-IgE Syndrome,[175-176] HIES) and cancers;[46] - In particular, the STAT3 transcription factor is of key importance to, and is over-expressed and constitutively activated in, a number of human cancers, including pancreatic cancer, melanoma, head and neck cancer, gastric cancer and breast cancer. The persistent activation of STAT3 plays a major role in up-regulating protooncogenes, resulting in survival and proliferation of cancer cells. Hence STAT3 is now a validated target for anticancer drug discovery.[22,52,177]

According to the original paradigm, cytokines (such as IL-6 and interferon) stimulate the phosphorylation of a specific tyrosine residue (Y705) in STAT3, which confers on it the ability to dimerize, subsequently translocate to the nucleus and bind to its consensus DNA sequences.[30] However, it has previously been reported that phosphorylation is not a pre-requisite for nuclear transport of STAT3 (which is dependent on *Ran* and importin α3 and β1), and a novel mechanism, where unphosphorylated STAT3 might have significant transcriptional control over the expression of genes that do not directly respond to phosphorylated STAT3 was suggested.[50,178] Furthermore, direct binding of the unphosphorylated STAT3 core directly to *M67* (i.e the high affinity STAT3 target DNA sequence) was recently reported.[179]

The transcriptionally active STAT3-STAT3 homodimer has been extensively targeted (both directly and indirectly)[49,51] by many research groups with the ultimate goal of suppressing the aberrant STAT3 function in human cancer cells. *In silico* screening techniques and structure-based drug design have been employed in a number of studies aimed at the discovery of small-molecule inhibitors of the STAT3-STAT3 dimerisation, since the prevention of dimerisation reflects directly into the downstream signalling thus avoiding blocking upstream signalling and reducing the side effect.

However, to date, only five categories of inhibitors have been reported.[63-67] This may be due to the inherent challenges of targeting protein-protein interactions (i.e lack of distinct binding sites); and in the case of inhibition of transcription factors, such as STAT3, the protein-DNA interaction may further increase the free energy of the associated protein-protein part, hence escalating the challenge of the latter's disruption by small molecules.[180]

There are currently two crystal structures of STAT3 deposited in the Protein Data Bank (PDB). Both these structures, of the phosphorylated STAT3β-DNA complex[181] (PDB id 1BG1) and the unphosphorylated STAT3 core fragment[182] (PDB id 3CWG), have regions of tertiary structure missing in the STAT3 protein-protein interaction domains. In particular, a flexible loop of residues containing the tyrosine residue Tyr 705, which is substantially enhancing the dimer stabilisation and the interaction surface, is missing from the structural data currently available. This flexible loop of residues that can reach across and interact with the partner SH2 domain, stabilise the association and bind the phosphorylated tyrosine into a specific binding site on the partner SH2 domain, and so uncertainty of the structural arrangement at this key region makes virtual screenings even more challenging. Furthermore, X-ray crystal structures are spatially and temporally-averaged, thus providing only limited information on dynamic properties.

In this part of the thesis, explicit solvent molecular dynamics simulations of the STAT3βtc homodimer:DNA complex will be addressed, in both its phosphorylated and unphosphorylated form, as well as in complex-bound and latent monomeric form, to gain a valuable insight into the aspects of the recognition at both protein-DNA (CHAPTER 4) and protein-protein (CHAPTER 5) level. These results will directly aid *in silico* approaches to the discovery of novel STAT3-STAT3 inhibitors for chemotherapeutic intervention.

CHAPTER 4:
Relationships between STAT3 mutations and protein-DNA recognition

4.1 BACKGROUND:

Recent molecular dynamics studies of the STAT3 complex[183] and the STAT3 SH2 domain with docked ligands[184] have focused on protein-protein interactions. However, understanding the protein-DNA recognition process is of importance for a complete insight into the dynamic nature of the STAT3 homodimer:DNA complex which is in turn critical for structure-based design of STAT3-STAT3 inhibitors. Water molecules often play a role in molecular recognition and association[185,186] and especially in protein stability,[187,188] protein-protein interactions, protein-ligand recognition,[185] and protein-DNA recognition.[189,190] Protein-DNA contacts at the atomistic level can be explained in terms of hydrogen bonds (direct readout), water-mediated hydrogen bonds, van der Waals, electrostatic and hydrophobic contacts.[190] Electrostatic charge interactions are considered to be a major determinant of protein-DNA interactions, contributing to indirect readout, depending on the distance between the charge groups. However, (weaker) hydrogen bonding contributes both to direct read-out (via contacts with base-pairs) as well as to indirect non-selective read-out (via phosphates and deoxyribose interactions). Therefore, if mutations occur at the protein-DNA interface, these interactions, and their interaction networks become affected and altered. Dominant negative mutations in the STAT3, mostly in the DNA-binding and SH2 domains, are associated with the Hyper-IgE Syndrome (HIES)[191] and these mutations have been mapped in STAT3 obtained from cells of clinical material.[175,176,191,192]

The present study examines the dynamic features of (1) the Tyr705-phosphorylated STAT3βtc homodimer:DNA complex, (2) unphosphorylated STAT3βtc monomer and (3) the 17 base-pair DNA with 5' overhanging ends, comprising the STAT3 binding site. The Tyr705 unphosphorylated STAT3βtc homodimer:DNA complex (4) is also examined, since evidence of unphosphorylated STAT3 binding to

DNA has been recently described by atomic force microscopy and by X-ray crystallography (Parkinson *et al* 2012, manuscript in preparation). A major focus is on the conformational changes at the protein-DNA interface with respect to STAT3-DNA complex formation, and the X-ray structure of the STAT3-DNA complex[181] (PDB id 1BG1).

4.2 AIMS:

Molecular dynamics (MD) simulations are used here to study the activated STAT3βtc homodimer:DNA complex, the latent unphosphorylated STAT3βtc monomer, and unphosphorylated STAT3βtc homodimer:DNA complex in an explicit water environment. The main questions set out to be answered by an analysis of the data obtained from MD simulations, each over a 50-ns time-frame are:

• How the transcription factor interacts with DNA, what is the nature of the conformational changes, and how does it compare to the dynamics of the unbound STAT3 monomer, and to the experimental (quasi-static) X-ray data?

• What are the key residues contributing to the recognition events involved in STAT3 protein-DNA interactions, and what role does the solvent play in the protein-DNA recognition?

• What are the specific protein-DNA contacts that are affected by the mutations in the DNA-binding domain, and what is their structural stability?

4.3 METHODS:

The three main stages of this study were:

(1) construction and preparation of the macromolecular systems

(2) calculation of explicit solvent classical MD trajectories

(3) data analysis

The MD trajectories were simultaneously used for generating multiple target conformations for a molecular docking study, described in CHAPTER 5.

4.3.1 Model building[2]

The STAT3 crystal structure[181] (PDB id 1BG1 at 2.25Å resolution) is deposited in the PDB as a monomer. The missing residues in the structure (185–193, 689–701, and 717-722) were identified as loops by the secondary structure prediction software JPRED[193] These were modeled by the ModLoop[194] program. The loops were joined to the main structure and subjected to energy minimization procedures using GROMACS[154] v 3.3.3 while keeping the position of the atoms of the core crystal structure fixed. The complete monomeric subunit was then used to construct the STAT3-DNA dimeric complex. The bases in the complementary DNA chain in the structure were modified using the Biopolymer software in the Insight II suite of programs (www.accelrys.com). The full model of the STAT3β homodimer:DNA complex was then subjected to a short cycle (1,000 steps) of molecular mechanics energy minimization to relieve any steric clashes from the structure.

[2] a model of the STAT3βtc homodimer:DNA complex was built by Dr. Shozeb Haider, before commencing my PhD studies in 2009.

4.3.2 System setup and molecular dynamics simulation

Each monomer of the generated STAT3 dimer comprised residues 136 to 716[3], while the DNA duplex, in a B-form conformation, comprised a 17-mer duplex with a nine base pair (9-bp) high-affinity binding site (*M67*) and overhanging 5' ends (Figure 4.1). This high-affinity M67 site d(TTCCCGTAA) differs at the 7th base from the consensus 9-bp target site d(TTCCCGGAA) previously described.[37] A model of the unphosphorylated STAT3-DNA complex was generated by removing the phosphate group of the phospho-Tyr705 (pY705) in the STAT3 dimer, and saving the complex as a new pdb file. A model of the unphosphorylated STAT3 monomer was generated by a corresponding manner, and saving the monomer as a new pdb file. The quality of the model was visually assessed by comparison and structural alignment with the crystal structure of the STAT3-DNA complex (PDB id 1BG1), and also with the crystal structure of the unphosphorylated STAT3 core fragment[182] (PDB id 3CWG). The calculated root mean-square deviation (RMSD) value of the corresponding protein residues of the unphosphorylated and phosphorylated STAT3 monomers was ~0.87 Å upon structural alignment using the PyMol[195] program (www.pymol.org) with the structures visually well matching (Figure 4.2).

[3] the truncated STAT3 protein **STAT3βtc** (residues 136-716) was used in the entire STAT3 *in silico* study.

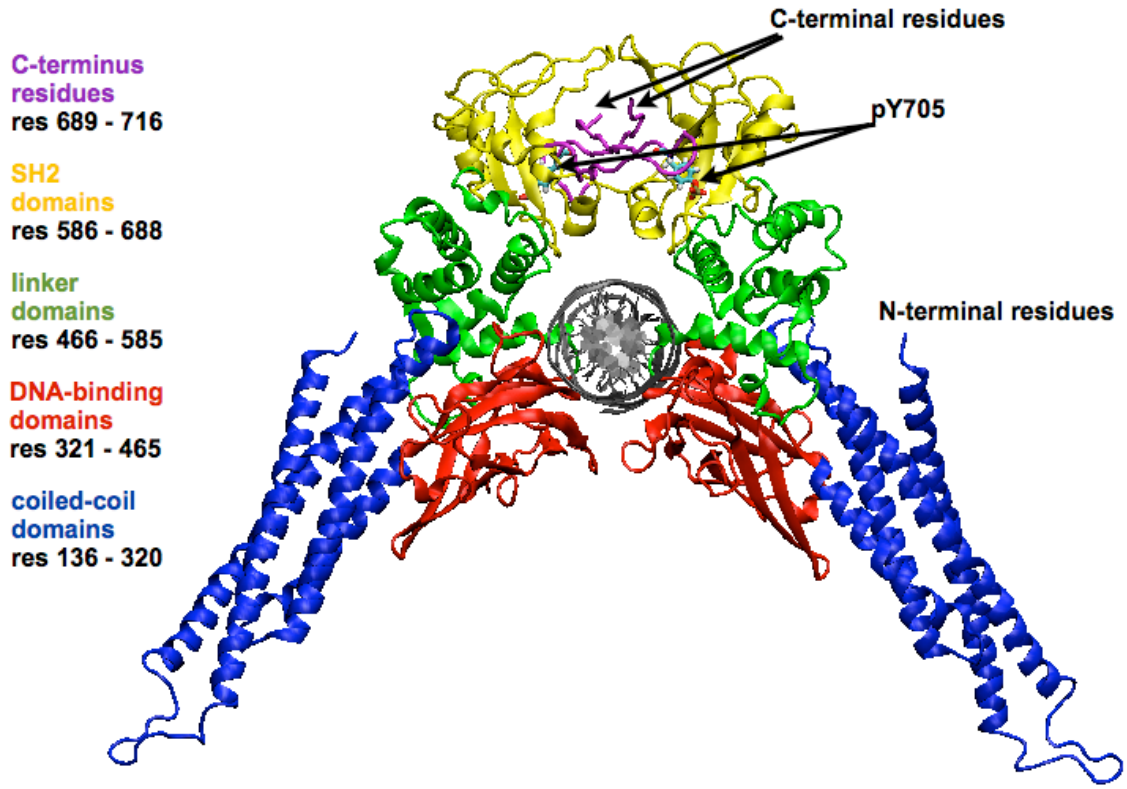


Figure 4.1: Model of the STAT3βtc-DNA complex.
Displayed in cartoon representation, individual domains are colour coded; Coiled-coil domains (*blue*), DNA-binding domains (*red*), linker domains (*green*), SH2-domains (yellow), and a stretch of ordered residues at the C-terminus (*magenta*), with the pY705 highlighted in stick representation (*cyan*). The 17-bp DNA with 5' overhanging ends is located between the two monomers (*grey*).

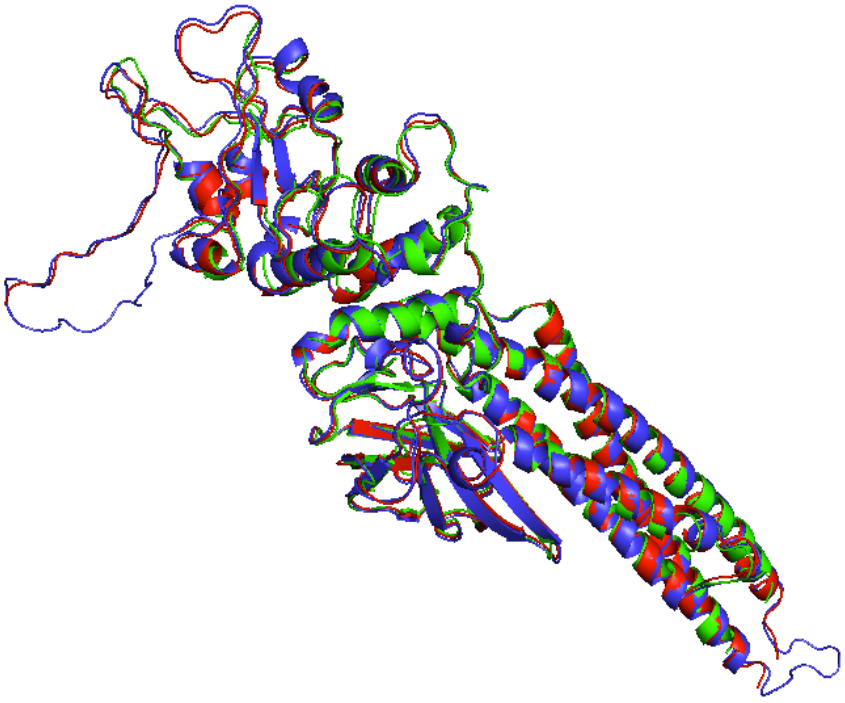


Figure 4.2: Structural alignment of the STAT3βtc monomers.
Unphosphorylated STAT3 core-fragment (PDB id 3CWG) (*green*); phosphorylated STAT3 monomer, used for model building (PDB id 1BG1) (*red*); and phosphorylated STAT3 monomer of the model (*blue*), all in cartoon representation.

All of the MD simulations were full-atom simulations and were performed with the GROMACS v 4.5.3[155] program, employing the improved protein side-chain torsion potentials from the AMBER parm99sb-ILDN[166] force field, together with the *parmbsc0*[161] force field, a refinement of the AMBER parm99[164] force field for nucleic acids. The *parmbsc0* force field was manually ported into GROMACS (as described in detail in CHAPTER 3). Further AMBER parameters for the phosphorylated-Tyr residues were also manually ported to GROMACS, by introducing a new phosphorylated-Tyr residue parameters[165] within the parm99sb-ILDN[166] force field. Details are given in the 'INTERMEZZO' section, CHAPTER 3. The modified parm99sb-ILDN force field port was then renamed and saved as a new entity in the working directory.

The simulation protocols were consistent for all four systems:
(1) phosphorylated STAT3βtc homodimer:DNA complex (pSTAT3-DNA complex)
(2) unphosphorylated STAT3βtc monomer (U-STAT3 monomer)
(3) 17-bp dsDNA STAT3 binding site (17-bp DNA)
(4) unphosphorylated STAT3βtc homodimer:DNA complex (uSTAT3-DNA complex)

The simulation systems were constructed by immersing the macromolecules in a dodecahedron of explicit TIP3P[196] water molecules, with a minimal clearance of 20 Å between periodic images for the starting configurations (i.e a diameter of the simulated system, plus twice the distance of 10 Å were applied to set the dimension of the dodecahedron box). Initially, hydrogen atoms were added to the simulated systems (except for *ds*DNA MD simulation) according to the protonation states of individual amino acids with protonable side chains at physiological pH. All Glu and Asp residues were unprotonated, hence assigned a negative partial charge (-1), while Lys and Arg residues were kept protonated (assigned with positive charge, +1). Cys residues were assigned as neutral, since STAT3 is a metal-free protein. Histidine side chains usually have pKa values close to physiological pH, and their protonation state determination can be a challenge. It has been reported[197] that His side-chains are protonated only if their side-chain nitrogen atoms are within 3.5 Å of a hydrogen bond acceptor in the structure, hence H437 was protonated in the model, while all other histidine residues were determined as neutral; in particular H147, H301, H332, H410 and H694 (Hδ) H447 and

H457 (Hε). Predicted pKa values were calculated by means of the PROPKA[198] web server. The negative net charge of the STAT3-DNA complexes, and the DNA duplex was neutralized by addition of excess positively-charged Na+ counter-ions, with negatively charged Cl- ions, providing a final NaCl concentration of ≈ 150 mM, which approximates physiological conditions. The counter-ions were automatically placed throughout the box by the *genion* utility of GROMACS 4.5.3 (www.gromacs.org) replacing solvent molecules, using the option to first position those atoms with the most favourable electrostatic potential.

The total number of atoms in the resulting solvated systems were 559,839 (pSTAT3-DNA complex), 371,422 (U-STAT3 monomer), 42,088 (17-bp DNA) and 559,814 (uSTAT3-DNA complex) respectively. Each of the systems was then allowed to adapt to the aqueous ionic environment (to remove any solvent-solvent and solvent-solute clashes created during the construction process) by applying 5,000 cycles of potential energy minimization, combining both steepest descent and conjugate gradient methods. This involved gradual relaxation of the initially-used harmonic restraints on the macromolecular atoms. With the greatest strain dissipated from the systems, the solvent was allowed to adapt to the macromolecules by being able to move freely, while keeping all atoms of the macromolecule harmonically restrained to their reference positions over the 150 ps period of molecular dynamics at 200K. Subsequently the unconstrained systems were slowly heated to 300K and equilibrated over 50 ps.

To compute MD trajectories, unrestrained production-level MD simulations were performed for 50 ns in the isothermal-isobaric ensemble (conserved NPT). Pressure and temperature were sustained at 1.0 bar and 300K, with temperature modulated by a velocity rescaling thermostat with a stochastic term,[199] and isotropic constant-pressure conditions controlled via the Parrinello-Rahman[200,201] algorithm. Non-bonded van der Waals interactions were calculated using Lennard-Jones 12-6 potentials with a 9 Å cut-off. Long-range electrostatics effects were calculated using the Particle-Mesh-Ewald algorithm (PME),[122] with a cut-off for the real-space term of 9.0 Å. The corresponding cut-off values of 9.0 Å for both non-bonded and long-range electrostatic interactions were chosen to fulfill the force field and PME[122] algorithm requirements of a

minimum cut-off value (< 8 Å), and to obtain superior performance with a minimum sacrifice in integration accuracy. The LINCS[126] algorithm was employed to constrain all bonds.[4] The integration time step applied was 2.0 fs with the coordinates saved every 5.0 ps. All MD simulations were computed on in-house Linux 64-bit Intel Core-i7 workstations, with efficient parallel scaling and double-precision calculations to prevent any energy conservation and stability issues. Trajectories were analyzed with the programs in the GROMACS 4.5.3 suite package, and visualized by means of the VMD[202] and PyMol[195] programs. All graphs were plotted using the Xmgrace program (plasma-gate.weizmann.ac.il/Grace).

4.3.3 Principal Component Analysis (PCA)

Principal component analysis (PCA, i.e. Essential Dynamics) enables large-scale correlated motions of atoms in the molecule to be identified and extracted from MD trajectories. Hence recurrent modes of structural changes can be eliminated from sets of structures, revealing the dominant modes and structures in the motion.[203] The backbone atom coordinates of the protein part and backbone atoms of the DNA helix of (1) the complete pSTAT3-DNA and uSTAT3-DNA complexes, (2) monomers-A, (3) monomers-B and (4) the U-STAT3 monomer were analyzed by PCA throughout the last 40 ns of the simulation time. PCA was then repeated considering only residues 321 to 688 in the protein part of the complex and the DNA helix, since the protein-DNA region is of primary interest in this study, and is stable throughout the MD simulation. The GROMACS program *g_covar* was used to calculate and diagonalize the mass-weighted covariance matrix. All structures from the MD trajectory were fitted to a reference structure representing the biomolecular system at the start of the production run. The generated eigenvectors, which provide a vectorial description of each component of the motion with their corresponding eigenvalues representing the energetic contribution of each component to the overall motion, were analyzed using the program *g_anaeig* in the GROMACS program suite. Protein and DNA backbone atoms only were used to construct the covariance matrix, since the size of the matrix varies with the square of the

[4] all-bonds constraints were applied, as a recommended setting at two different GROMACS workshops attended between 2009-2010.

number of atoms for which the covariance is calculated and Cα atoms alone may adequately sample large-scale correlated motions in proteins.[204] The conformational changes of the STAT3 biomolecular systems were subsequently visualized by means of the porcupine plots,[205] which were used to demonstrate the direction and magnitude of the extremes of fluctuations which were accounted for in the first and second eigenvector. The length and direction of the porcupine "needles" clearly indicates the scope of that motion. The porcupine plots were generated using a porcupine script[88] and visualized using the VMD[202] program.

4.3.4 Cluster Analysis

By definition, cluster analysis is designed to detect hidden clusters in a set of objects which are described by numerical or structural data so that the members of each cluster behave similarly to each other and groups are well separated.[206] To identify clusters of structures in a trajectory, the RMSD can be used to assign distances between cluster sets with respect to the distances between structures, reflecting the range of conformations and their relative populations. The *gromos* agglomerative clustering algorithm[207] was implemented via the GROMACS clustering utility (*g_cluster*). This was employed to extract the clusters of conformers in the STAT3β-DNA complexes generated over the simulation time-frame, with solely the duplex DNA, DNA binding domains, linker and SH2 domains (i.e residues 321 - 688) being considered, for the same rationale as described above. The RMSD cut-off distance was 2.0 Å for two structures to be considered as neighbors. The initial 2 ns of the MD trajectories were rejected, and the last 48 ns were used for the analysis.

4.3.5 Protein-DNA and water contact-residues analysis

Structures representing the conformations of the pSTAT3-DNA and uSTAT3-DNA complex generated over the course of the simulations obtained from the cluster analysis were subjected to contact analysis, focussing on the protein-DNA interface and water contacts. Interfacial interactions were defined on the basis of physicochemical and distance criteria between atoms. A 3.2 Å donor-acceptor distance for hydrogen bonds, and 5.0 Å for electrostatic interactions between the charged side-chains of residues, and hydrophobic interactions was used. From the MD simulation generated trajectories, the interatomic distances between the hydrogen bond-forming residues were calculated using the GROMACS program *g_mindist*, and the overall time spent within the specified distance for each hydrogen bond was determined in terms of percentage of existence of the individual hydrogen bonds over the course of the simulation (i.e 48 ns that were analyzed). Also all the solvent molecules were analyzed to determine their interactions with the protein-DNA complex in terms of time spent within the hydrogen bonds formed with the protein, DNA or both (bridging waters). Water molecules were only accepted if they formed hydrogen bonds for longer than 1 ns. Dedicated tools in the Chimera[208] and VMD[202] programs were used for hydrogen-bond detection and analysis.

4.3.6 Water density maps[5]

Water density maps were generated from snapshots (100 frames, 500 ps time-frame) of the protein-DNA complex in explicit water that were generated throughout the simulation. For each area of interest, three consecutive residues on the same chain were chosen and used as a reference set to align all of the snapshots, using the program LSQMAN.[209] The water coordinates were extracted from each of the aligned files and placed in a single file. These water files were then converted into density maps (.mtz file extension) using the CCP4 programs SFALL and FFT.[210] Maps were then visualized with the PyMol[195] program, by means of the *density map wizard* suite.

[5] The structural alignment followed by data extraction for water density maps was performed by Dr. Alan K. Todd

4.4 RESULTS AND DISCUSSION

This study has focussed on four main areas:

(1) nanosecond-scale trajectory analysis in terms of structural stability and conformational variability
(2) principal component analysis revealing the nature of the major concerted motions
(3) cluster analysis providing structures for mapping both the protein-DNA, and protein-protein interface (the latter being a focus of the following CHAPTER 5) together with an analysis of the hydration at the protein:DNA interface
(4) detailed mapping of the mutations in the DNA-binding region and their location to specific contacts

The unphosphorylated unbound (U-STAT3) vs. the phosphorylated complex-bound STAT3 monomers are also compared (Figure 4.2; shown in the methods section), together with the effects of STAT3 dimerization followed by STAT3β-DNA complex formation with the DNA duplex target sequence. The unphosphorylated STAT3β-DNA complex is also brought into perspective to complement the study. These findings are then related to the X-ray crystal structure[181] (PDB id 1BG1) that was used for the model building. In general the simulation results confirm earlier observations from the crystal structure, but are able to extend them to provide a dynamic picture of this protein:DNA system.

A view of the pSTAT3β-DNA complex is shown in Figure 4.1 (methods section). Each of the two monomers forming the anti-parallel dimer, is composed of four domains: (1) a coiled-coil domain formed by a N-terminal four-helix bundle (residues 136 to 320), (2) a DNA-binding domain comprising an eight-stranded β-barrel (residues 321 to 465), (3) an α-helical linker domain (residues 466 to 585), (4) a SH2 (Src Homology 2 domain (residues 586 to 688) with a stretch of ordered residues at the C-terminus, (residues 689 to 716) containing the important pY705.[181]

4.4.1 Structural stability and conformational variability

The overall 200 ns simulation for the four STAT3 models resulted in energetically conserved and stable simulations for the ~ 560,000-atom STAT3-DNA complexes, ~ 372,000-atom U-STAT3 monomer, and ~ 42,000-atom 17-bp DNA target sequence systems (Figure 4.3).

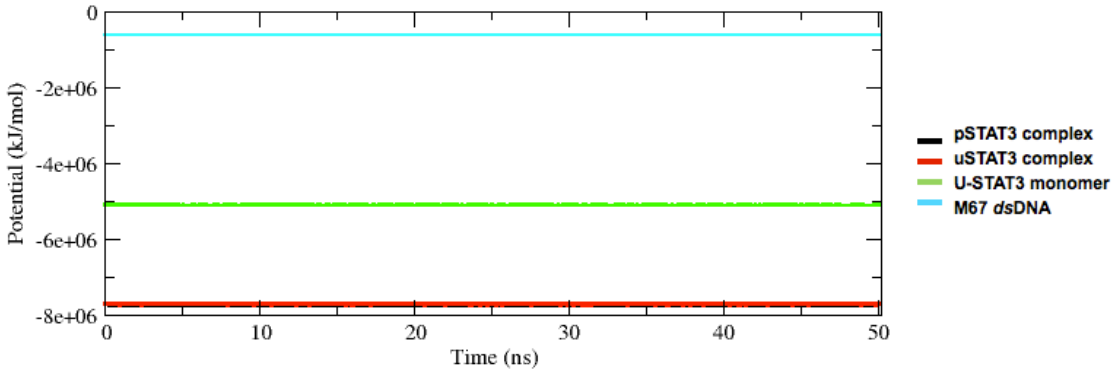


Figure 4.3: Potential energy of the simulated systems as a function of time.
Plot is shown for the pSTAT3-DNA complex (*black*), uSTAT3-DNA complex (*red*), U-STAT3 monomer (*green*), and 17-bp DNA (comprising the STAT3 binding-site) (*cyan*).

RMSD values for the backbone atoms as a function of the simulation time were used as a measure of stabilization of the four models during the simulation, comparing both initial reference, and time-averaged structures of the models, with the latter providing superior insight into the structures reaching the plateau (Table 4.1 and Figure 4.4).

Table 4.1: Summary of the simulations together with the numbers of residues and RMSD values.

				Environment		
Simulation system		**Timescale**	**Residues**	**Water TIP3P**	**RMSD bb vs. start**	**RMSD bb vs. avg**
pSTAT3-DNA complex	(1	50 ns	1,198	179,628	3.92 Å	2.05 Å
pSTAT3 monomer-A			581		2.97 Å	1.93 Å
pSTAT3 monomer-B			581		3.37 Å	1.64 Å
17-bp DNA			36		2.46 Å	1.35 Å
U-STAT3 monomer	(2	50 ns	581	120,459	5.19 Å	2.25 Å
free 17-bp DNA	(3	50 ns	36	13,695	4.05 Å	2.32 Å
uSTAT3-DNA complex	(4	50 ns	1,198	179,623	3.80 Å	2.22 Å
uSTAT3 monomer-A			581		3.45 Å	1.69 Å
uSTAT3 monomer-B			581		3.27 Å	1.70 Å
17-bp DNA			36		2.46 Å	1.51 Å

The conformational stability of the higher-order structures was also examined over the course of the simulation. The trajectory of the pSTAT3 complex was stabilized at around ~ 3.9 Å (and at around ~ 2.0 Å with respect to the time-averaged structure obtained from the MD), with the individual corresponding monomers-A and -B being stabilized with ~ 0.5 Å difference in their respective RMSDs. Similarly, the uSTAT3 complex was stabilized at around ~ 3.8 Å (and at around ~ 2.2 Å for the time-averaged structure), with the individual uSTAT3 complex-bound monomers-A and -B being stabilized with ~ 0.2 Å difference in their respective RMSDs. Whereas there is a modest difference in the stability of the pSTAT3 complex-forming monomers, the uSTAT3 complex-bound monomers show very little difference in their RMSDs with respect to both their initial, and time-averaged structures. However, in both cases, the complex-bound monomers show improved stability comparing to the unphosphorylated STAT3 monomer. The slightly larger fluctuations in the U-STAT3 model suggest a stabilizing effect of the STAT3β-DNA complex formation towards both the STAT3 monomers and the DNA duplex (Table 4.1, Figure 4.4).

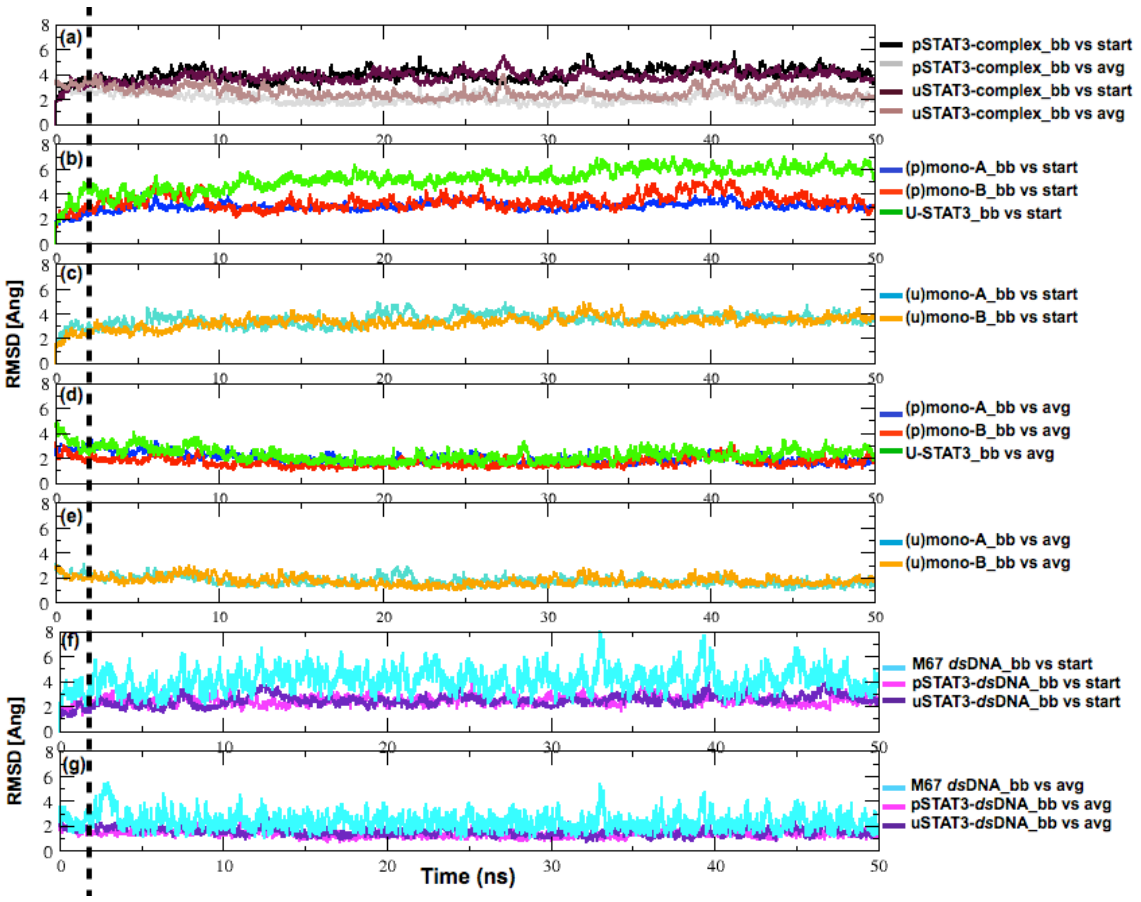


Figure 4.4: RMSD plots following the stability of the STAT3 models.
RMSD plots of the backbone atoms for the STAT3β-DNA complexes versus the starting (*black/maroon*) and time-averaged (*grey/brown*) structure (a), for the individual STAT3β-DNA complex-forming monomers-A (*blue/turquoise*), and monomers-B (*red/orange*), unbound unphosphorylated monomer (*green*) versus the initial (b, c) and time-averaged (d, e) structures; and for the complex-bound 17-bp dsDNA (*magenta/indigo*), and 17-bp M67 dsDNA (*cyan*) versus the initial (d) and time-averaged structures (e).

The flexible regions in the protein part of the STAT3 complexes, and the U-STAT3 monomer were also analyzed by examining their structural fluctuations in terms of root mean square fluctuations (RMSF) as a function of residue number (Figure 4.5). Large fluctuations indicated by sharp peaks correspond to the loops. The magnitude of fluctuations is most significant at both terminal regions of the models and in the DNA-binding domain, indicating significant interaction with the duplex DNA, in accord with the crystal structure. Differences in the fluctuations within individual domains were observed for monomers A and B in both pSTAT3 and uSTAT3 complexes (Figure 4.5 a,c) so that despite the identity of their primary sequence, there is a real difference in their dynamic behavior.

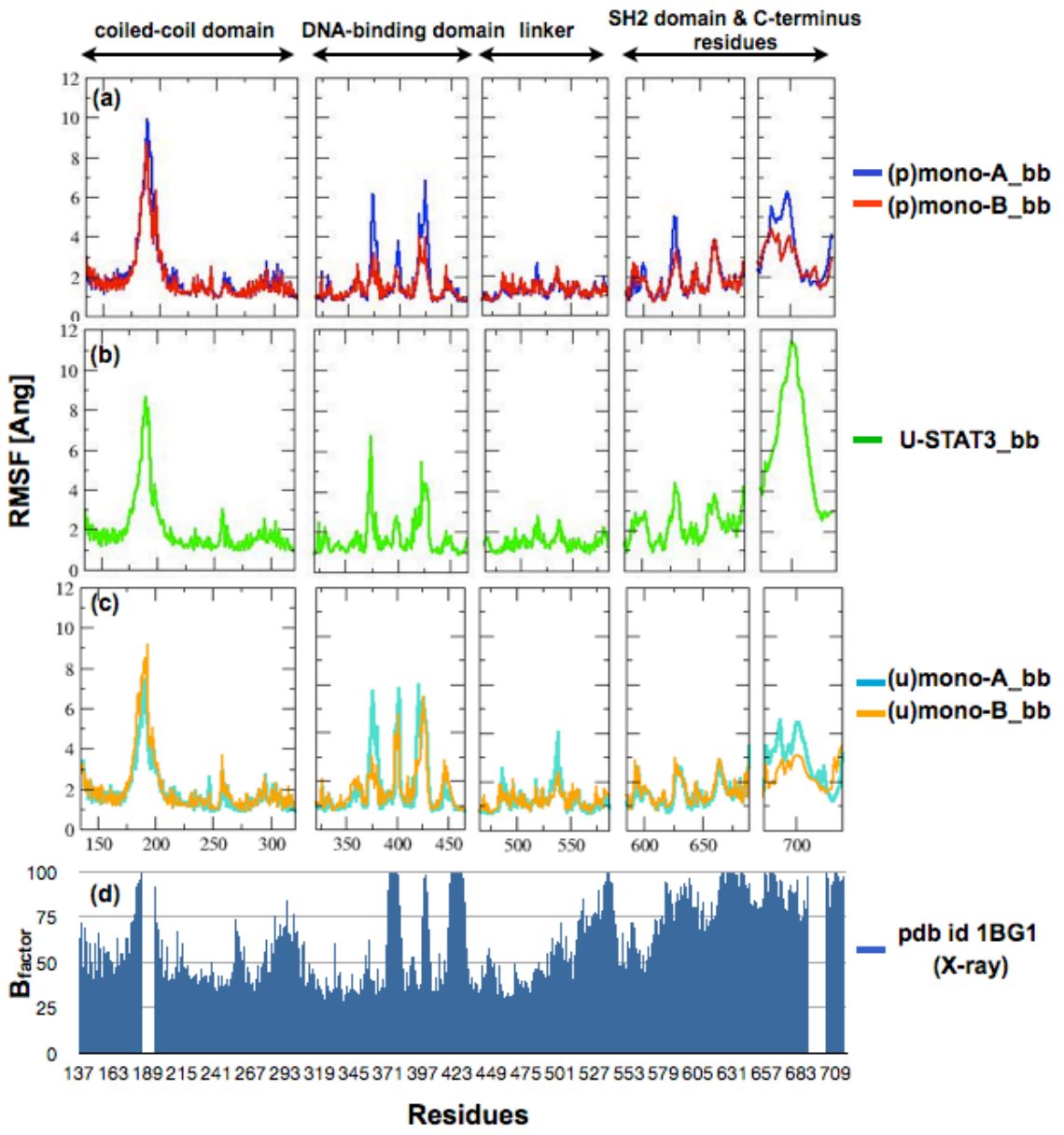


Figure 4.5: RMSF plots of the STAT3 models in comparison with experimental B-factors.
RMS fluctuations averaged over each amino-acid residue for (a) pSTAT3β-DNA and (c) uSTAT3-DNA complex forming monomer-A (*blue/turquoise*) and monomer-B (*red/orange*), and for (b) unphosphorylated U-STAT3β monomer (*green*) during the dynamics runs. (d) The RMS fluctuations of the STAT3β models may be compared with the experimental B-factors for the STAT3β crystal structure (PDB id 1BG1), which was used to build the initial model.

The complimentary pattern of fluctuations can be observed for the U-STAT3 monomer with the largest difference at the SH2 domain and in particular in the elongated stretch of C-terminus residues (Figure 4.5 b). Whereas in the pSTAT3/uSTAT3 complex-bound form the "elongated arm-like" loop-forming residues 689-716 interact with the other STAT3 monomer (therefore the magnitude of fluctuations is partially limited), in the

case of the U-STAT3 monomer the terminal loop was observed to be free to move and fold onto the SH2 domain which may explain the very large RMSF values for the U-STAT3 monomer C-terminal region. In terms of the pSTAT3 and uSTAT3 complexes, larger fluctuations of the DNA-binding domains and linker domains within the uSTAT3 complex can be observed with respect to pSTAT3 (Figure 4.5 a,c), whereas the protein-protein interacting domains (i.e SH2 domains with the stretch of C-terminal residues) of the pSTAT3 complex display larger fluctuations comparing to the uSTAT3 model. The pattern of experimental B-factor values in the crystal structure 1BG1, plotted as a function of residue number (Figure 4.5 d) show certain similarities to the fluctuations observed in the STAT3 MD simulations. However, due to the extremely large values of the B-factors for atoms in the SH2 domain and the uncertainties surrounding the stretch of C-terminal residues in particular (with a number of residues missing), the STAT3 conformations generated here during the MD run can provide greater insight into the arrangement of the protein-protein interaction than the crystal structure. A comparison has also not been made between the crystallographic water molecules and those in the MD simulations in view of the uncertainties surrounding the former, which also mostly have high B-factors. For any further analysis, the first 2 ns of the simulations were rejected, as a consequence of relaxation of the system, after which the trajectory of the STAT3 systems was stable (Figure 4.4).

Since the major focus of this study is on the conformational changes at the protein-DNA interface, with respect to the pSTAT3-DNA complex formation and to the X-ray structural data of the STAT3-DNA complex, RMSD plots on a per residue basis were generated for the DNA-binding domain, linker and SH2 domain (residues 321 to 688) of both pSTAT3 complex-bound monomers-A and -B, as well as for the U-STAT3 monomer, in order to obtain more insight into the stability of this key region (Figure 4.6). These domains of monomer A and B were observed in the simulations to have comparable stabilities. Thus the mean RMSD per residue values are: DNAbin domain of monoA: ~ 2.8 Å, monoB: ~ 3.3 Å, Umono: ~ 3.0 Å; linker of monoA: ~ 2.1 Å, monoB: ~ 2.6 Å, Umono: ~ 2.4 Å; SH2 domain of monoA: ~ 3.5 Å, monoB: ~ 4.3 Å, Umono: ~ 4.1 Å for the U-STAT3 monomer domains. The residues determined at the protein-DNA interface are highlighted (in light-grey) and the residues carrying the

mutations are marked in Figure 4.6, which will be discussed in more detail in the following sections.

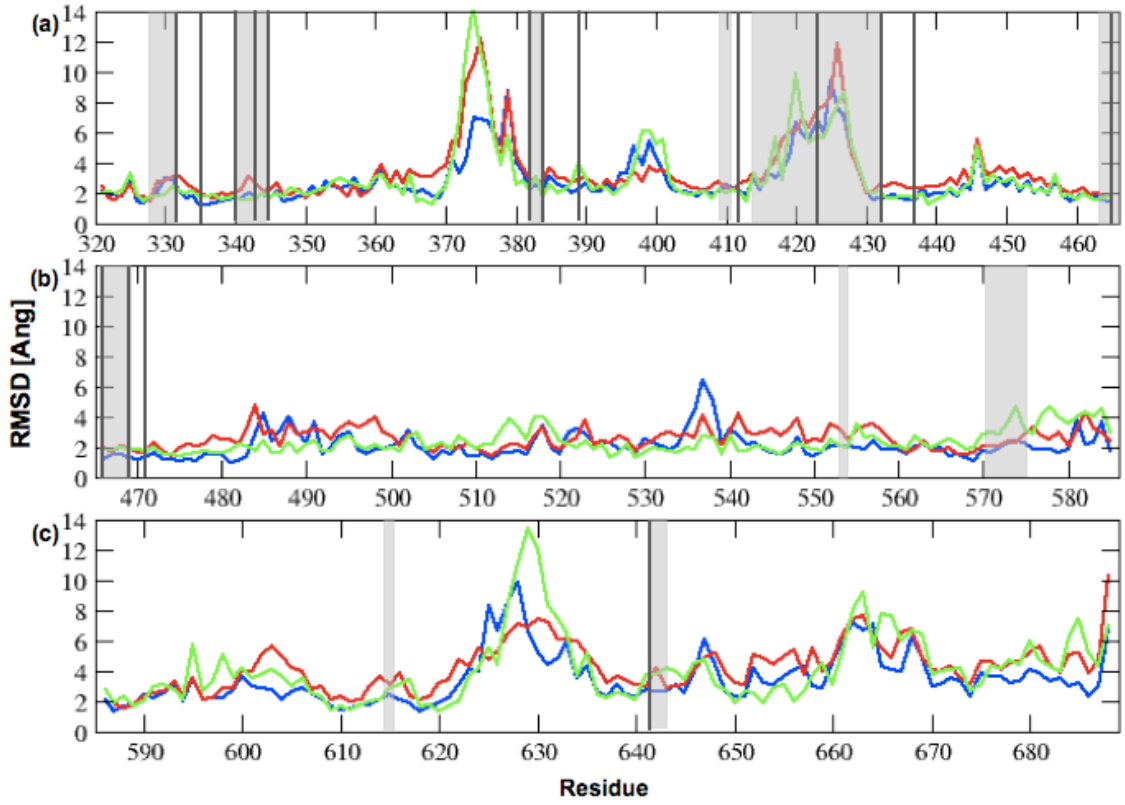


Figure 4.6: RMSD per residue basis for the STAT3 core region.
(a) DNA-binding domain (residues 321-465), (b) linker (residues 466-585) and the (c) SH2 domain (residues 586-688) of the individual pSTAT3-DNA complex-forming monomer-A (*blue*), and monomer-B (*red*) unbound unphosphorylated monomer (*green*). Residues determined at the protein-DNA interface are highlighted in light-grey and the residues reported to be carrying the mutations are marked in by grey line.

Backbone-atom RMSD plots of the 17-bp duplex DNA (Figure 4.4 f,g) unsurprisingly show that STAT3-DNA complex formation has an overall stabilizing effect on the DNA, in agreement with the RMSF plots of the complex-bound and free 17-bp *ds*DNA (Figure 4.7), with the pSTAT3-DNA complex showing even stronger effect on the DNA stabilization, resulting in overall smaller fluctuation of the DNA basis throughout the MD run. The terminal regions display larger fluctuations comparing to the rest of the DNA helix since both the 5' ends are overhanging. Furthermore, the amplified fluctuation at the T5' region in the STAT3-DNA complex-bound forms can be explained by a strong interaction of the residues in monomer-A with the C(8) and G(7) nucleotides of the

opposite DNA chain, hence unwinding the duplex DNA at that terminal. The protein-DNA interactions will be discussed in more detail in the following section.

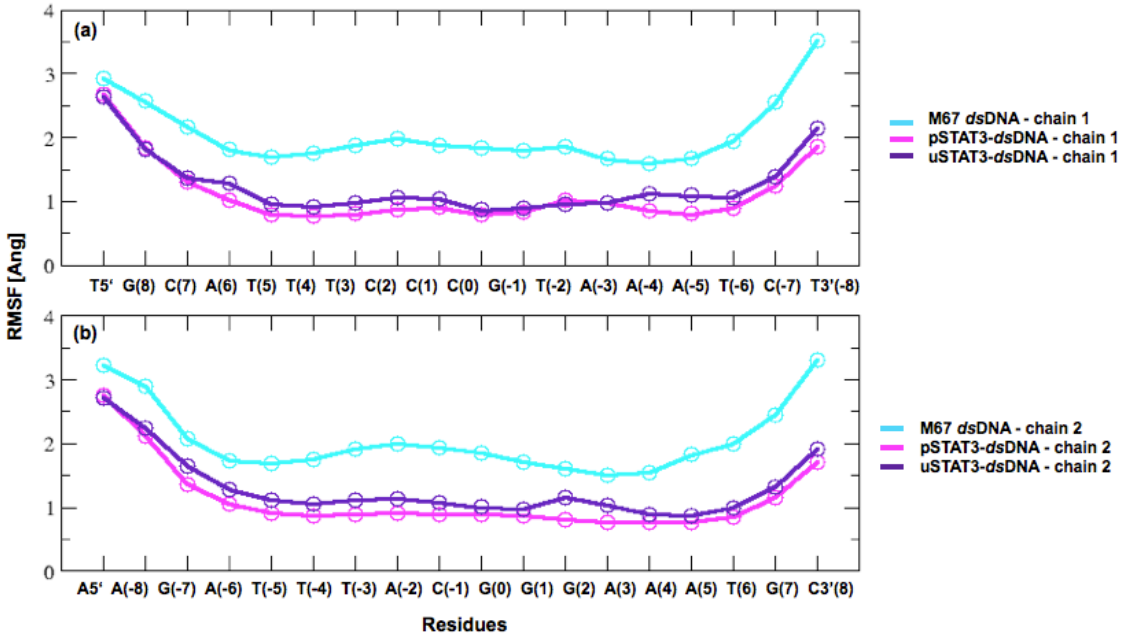


Figure 4.7: RMSF plots for the 17-bp *ds*DNA consensus sequence.
RMS fluctuation plots for the model of the 'free' 17-bp *ds*DNA duplex (*cyan*), pSTAT3 complex-bound DNA duplex (*magenta*), and uSTAT3 complex bound *ds*DNA (*purple*), all containing the *M67* consensus sequence; the corresponding DNA chains are plotted together: (a) chain 1 and (b) chain 2.

4.4.2 PCA: Defining the concerted motions in STAT3

The overall patterns of motions in both phosphorylated and unphosphorylated STAT3β homodimer:DNA complexes (and their individual monomers), and the U-STAT3 monomer, were identified via principal component analysis (PCA), a method that extracts the dominant modes in the motion of the molecule from the trajectory obtained from the MD simulation, and provides a quantitative assessment of the correspondence between the MD data sets. PCA was performed on the backbone atoms of the STAT3 models, employing the trajectory from the last 40 ns of each of the simulations (with a 25 ps time-step). The total variance was accounted for by the calculate eigenvectors and the corresponding eigenvalues from the covariance matrix of simulation. Two types of PCA were carried out: (1) including all backbone atoms of the protein-DNA complexes

(i.e residues 136 to 716 and *ds*DNA if relevant), outcome of which was further subjected to graphical visualization by means of the porcupine plot as described below; and (2) considering only residues 321 to 688 of the protein with the DNA duplex (i.e DNA binding domains, linkers and SH2 domains). The first 10 eigenvectors were considered for further analysis, which showed that the first three eigenvectors account for ~50% of all the motion in the simulated systems (Figure 4.8).

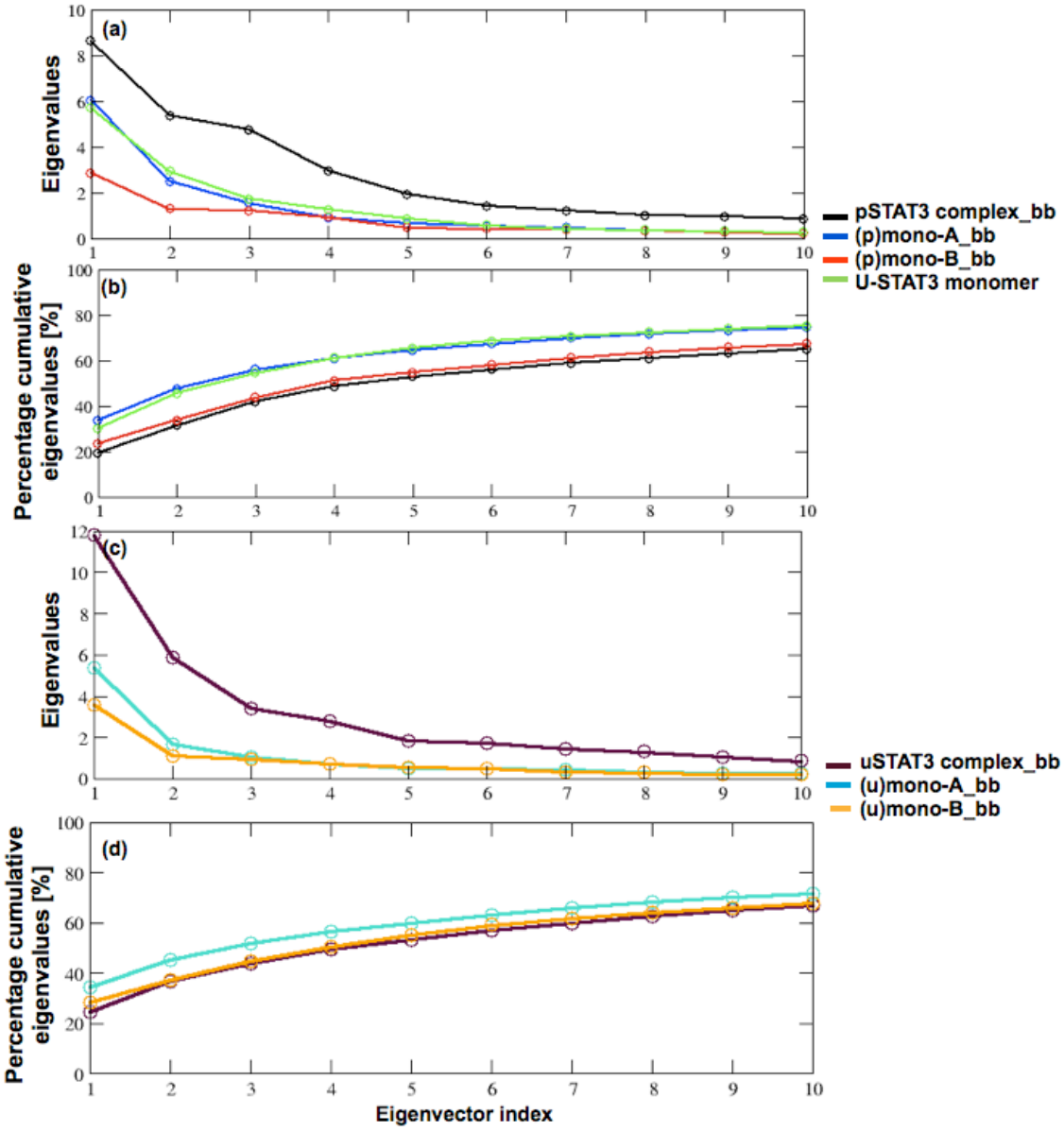


Figure 4.8: PCA graphs capturing the concerted motions of the STAT3 models.
(a, c) Eigenvalue versus eigenvector index, and (b, d) cumulative eigenvalue (%) derived from PCA of backbone-atom MD trajectories for the first ten eigenvectors of the (a, b) pSTAT3-DNA complex (*black*), for the individual complex bound monomer-A (*blue*), and monomer-B (*red*), and U-STAT3 monomer (*green*); and for the (c, d) uSTAT3-DNA complex (*maroon*) and its individual complex bound monomer-A (*turquoise*) and monomer-B (*orange*). Only protein residues 321- 688 and DNA helix are considered.

The first two eigenvectors were represented by means of porcupine plots that show the scope and magnitude of the motion (given by the first and second eigenvector respectively) for each of the backbone atoms, for residues 136-716 and DNA helix (Figure 4.9), for both pSTAT3-DNA complex and uSTAT3-DNA complex. The orientation is such, that complex-bound monomer-A is always on the right within the Figure 4.9. The complex-bound motion of STAT3 monomer-A is compared with the U-STAT3 monomer (Figure 4.9 g, h). The first principal component for the STAT3-DNA complex confirms the observation made from the RMS plots that the most prominent motions are observed at the loops within the coiled-coil domain (the loop connecting helices α1 and α2), resulting in a characteristic scissor-like motion, with the duplex DNA in the middle acting as a “hinge”.

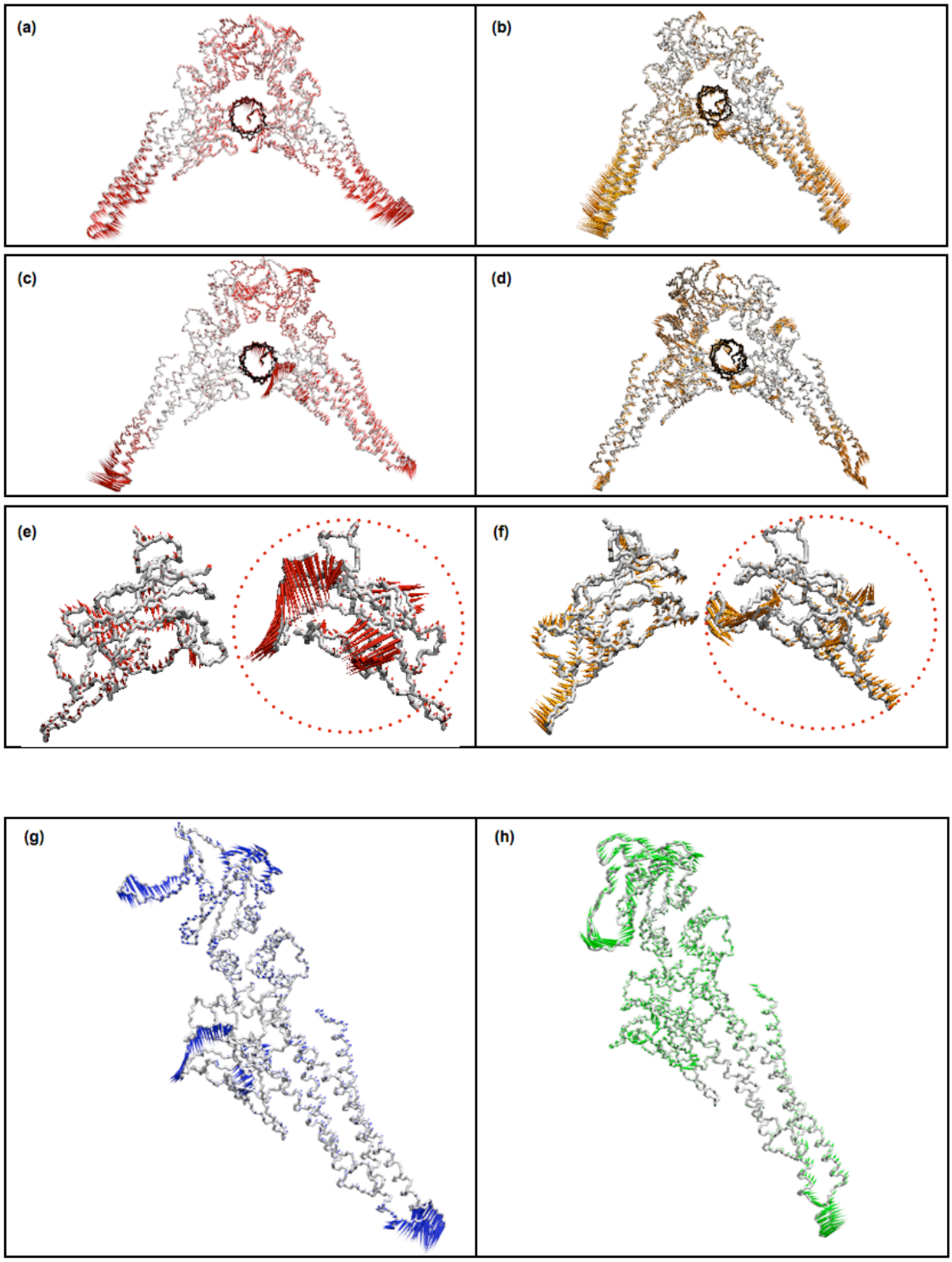


Figure 4.9: Porcupine plots for the STAT3-DNA complexes and STAT3 monomer.
Porcupine plots for the (a) first and (c) second eigenvector for the pSTAT3 complex simulation with detailed focus on the characteristic movement of the monomer-A DNA-binding domain at the protein-DNA interface, which differs significantly from that of the monomer-B DNA-binding domain movement (e). Complementary porcupine plots of the (b) first and (d) second eigenvector for the uSTAT3 complex and the close-up look at the DNA binding domains motion (f). DNA-binding domain of the monomer-A is in both cases marked in red-dotted circle; (g) porcupine plots of the first eigenvector for the pSTAT3 complex-bound monomer-A (*blue*), and (h) unphosphorylated unbound U-STAT3 monomer (*green*). The models are shown as bond trace with the arrows attached to the individual backbone atoms representing the scope and the magnitude of the motion.

Several observations can be made from the porcupine plots (corresponding to data from the RMS plots):

(1) monomer-A undergoes overall greater motion than monomer-B; however the difference is smaller in the case of the uSTAT3-DNA complex.
(2) the dynamics/motion of the DNA-binding domain of monomer-A suggests that the DNA-interacting loops of monomer-A are inserted deeper into the DNA groove. This results in a greater span of interaction with the DNA duplex (further discussed below), and also indicates DNA unwinding at the T5' end. The larger magnitude of movement at the T5' of the DNA is also in accord with this observation, as shown in Figure 4.9. Similar patterns of movement, with monomer-A inserting further into the groove of the DNA helix, was observed within alternative MD simulations employing the ff03 force field, cubic solvation box and GROMACS v 3.3 (data not discussed). This then suggest that the described difference in STAT3 monomers movements was not an artifact of the simulation.
(3) when comparing the DNA-binding domains of the pSTAT3 and uSTAT3 complexes, DNA-binding domain of monomer-A of the pSTAT3-DNA complex undergoes significantly larger motion, as shown in Figure 4.9 (e, f)
(4) the major flexibility at the SH2 domains region arises from the loop-like stretch of C-terminal residues (residues 689-716) that forms an arm which reciprocally binds onto the other monomer.

With respect to the predominant motion within the U-STAT3 monomer, the overall magnitude of fluctuations at the incremented regions (protein-DNA interface) is smaller comparing to those for monomer-A (Figure 4.9 g, h), since there are no interaction partners. The most significant difference in the movement lies within the SH2 domain with the folded C-terminal stretch of residues, and the direction of the movement is opposite compared to that in complex-bound monomer-A.

4.4.3 Cluster analysis: Statistical description of the interface

Cluster analysis was applied to the large amount of data generated in order to provide a statistical description of the dynamics of the STAT3-DNA complexes. Here, only residues 321 to 688 of both monomers (i.e DNA-binding domains, linker domains and SH2 domains) together with the DNA helix were used for clustering. This analysis used the 1920 frames extracted from the MD trajectory at a time interval of 25 ps (i.e 48 ns) for the matrix construction, with a 2.0 Å RMSD cut-off applied for the neighbor search. The first five of the 12 resulting clusters of the pSTAT3 protein-DNA complex are represented by middle structures (Figure 4.10) – these are the conformations sampled during the simulation at 300K. These were used for the analysis of the protein-DNA interface detailed below. The first five clusters covered approximately 98% of the total ensemble of sampled conformational space. Cluster 1 was the predominant conformer, being populated approximately 57% of the simulation time, whereas cluster 2 spanned over 10 ns of the simulation (~ 22% of the simulation time). In terms of the uSTAT3 protein-DNA complex, 16 clusters were obtained from the corresponding trajectory (48 ns, 1920 frames with 25 ps time steps). The first five of the 16 resulting clusters, represented by middle structures, spanned through approximately 90% of the total ensemble of sampled conformational space at 300K. Cluster 1 was populated approximately 46% of the simulation time, being the predominant conformer, while cluster 2 represented nearly 10 ns of the simulation (~ 17 % of the simulation time). These five middle structures from the first five clusters were also subjected to the protein-DNA interface analysis, hence provide a useful comparison with the phosphorylated STAT3 protein-DNA interface. Larger number of clusters of the uSTAT3 complex indicates, that despite good overall stability, there are indeed increased fluctuations in this modeled system resulting in more representative structures than for the pSTAT3 complex. Since the only difference between the two complexes is the phosphorylation of the specific amino acid residue, Y705, this may further indicate the increased association strength of the pSTAT3-DNA complex, which is in accord with the experimental observations. The characteristic features of the phosphorylated Y705 interaction with the partner SH2 domain are fully described in CHAPTER 5, employing the exact five clusters of the pSTAT3 protein-DNA complex.

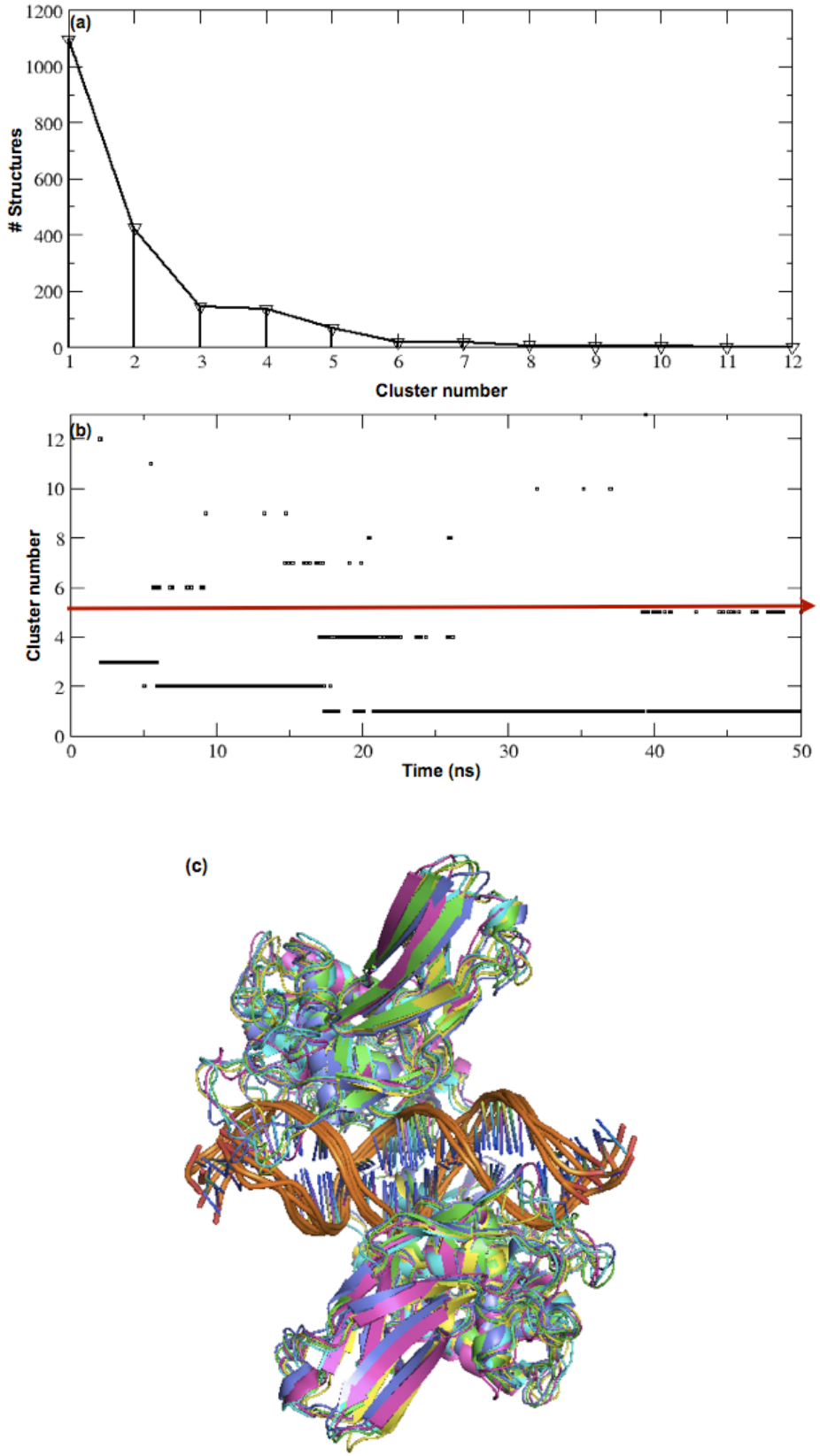


Figure 4.10: Cluster analysis of the DNA-binding region of the pSTAT3 model.
Structures representing the pSTAT3-DNA complex (residues 321-688 of both monomers and the DNA duplex), obtained from the trajectory over the dynamics and gathered into 12 clusters of decreasing sizes (a), plotted against the simulation time (b). Only the middle structures of the first 5 clusters were used for the protein-DNA interfacial analysis (c).

Representative structures of the true conformational space at 300K obtained from the cluster analysis of the 25-ns MD trajectories (2.5 Å cutoff, 1000 frames) subsequently provided multiple target conformations for a small molecule docking study, which is reported in CHAPTER 5. This approach is supported by recent studies demonstrating that state-of-the-art docking algorithms predict an incorrect binding pose for about 50-70% of ligands when only a single fixed receptor conformation is considered.[141]

4.4.4 Mapping the protein-DNA-solvent interaction

The middle structures of the first five clusters obtained from the cluster analysis were used to describe the protein-DNA interactions, as they represent over 98% (90% for the uSTAT3 complex respectively) of the conformational space covered throughout the 50-ns MD run. Within a 5.0 Å distance cut-off, and a condition of at least one non-hydrogen atom present in that distance range, 25 corresponding residues were identified at the protein-DNA interface of the pSTAT3β-DNA complex, within each of the subunits (monomer-A and -B). A further seven non-corresponding residues of monomer-A (G419, N420, G421, G422, A424, N425 and C426) and seven non-corresponding residues of monomer-B (M329, P330, H410, N553, K573, E616 and K642) were observed at the interface (Table 4.2[6]). Although the monomers are identical in terms of the reciprocal pY705 interaction at the SH2 domains and the high-affinity binding predominantly in the DNA major groove, the partial asymmetry of the monomeric protein-DNA interface and the specific dynamics of each monomer may be explicable by the asymmetric spiral (or "screw-like") shape of the DNA duplex. A range of electrostatic, hydrophobic, and hydrogen bonds-forming contacts between the STAT3 protein and the DNA duplex occur predominantly in the DNA-binding domains, although they also occur in the linker domains of both phosphorylated STAT3β monomers, and at the SH2 domain of monomer-B (E616, K642 and Q643).

[6] * There are 39 residues in the interface, 25 of which are common to the two monomers. Residues forming hydrogen bonds with the DNA are highlighted in grey, and unless stated otherwise, the hydrogen bonds involve the DNA backbone atoms. Hydrogen bonds comprising the atoms of the DNA bases are listed. Point mutations at the protein-DNA interface are shown in dark grey, and residues forming hydrogen bonds with the solvent molecules are marked (WAT), with water molecules bridging the protein-DNA interaction indicated as (WAT-bridge).

Table 4.2: Contact residues within the pSTAT3 protein-DNA interface.
Amino acid residues of the pSTAT3 complex-bound monomers-A and -B determined to be within the 5 Å cut-off protein-DNA interface of the first 5 middle structures obtained from the cluster analysis*.

monomer-A					monomer-B			
HB -solvent	**Occup.**	**HB-DNA**	**5Å int-Face**		**5Å int-Face**	**HB-DNA**	**Occup.**	**HB-solvent**
			--	**M 329**	✓			
			--	**P 330**	✓			
WAT			✓	**M 331**	✓	**dC(0)**	95%	
WAT			✓	**H 332**	✓	**dC(0)**	98%	
	60%	**dG(0)**	✓	**K 340**	✓	**dG(-1)_O6**	24%	3x WAT-bridge
WAT-bridge			✓	**T 341**	✓			2x WAT-bridge
			✓	**G 342**	✓			
			✓	**V 343**	✓			WAT-bridge
2x WAT-bridge	22%	**dA(-2)**	✓	**Q 344**	✓	**dC(+2)**	53%	
	73%	**dC(-1)**				**dC(+1)**	33%	
WAT-bridge, WAT	98%	**dT(+4)**	✓	**R 382**	✓	**dT(-4)**	97%	
			--	**H 410**	✓			
	2%	**dT(+3)**	✓	**R 414**	✓			
4x WAT-bridge, 5x WAT			✓	**E 415**	✓			5x WAT-bridge, 3x WAT
	69%	**dT(+4)**	✓	**R 417**	✓	**dT(-4)**	57%	
	83%	**dT(+3)**				**dT(-3)**	91%	
			✓	**G 419**	--			
	4%	**dG(+7)**	✓	**N 420**	--			
	7%	**dC3'(+8)**						
	8%	**dC3'(+8)**	✓	**G 421**	--			
	11%	**dC3'(+8)**	✓	**G 422**	--			
WAT	8%	**dA(+6)**	✓	**R 423**	✓	**dT3'(-8)**	5%	
	10%	**dT(+5)**				**dT(-5)**	42%	
	27%	**dC3'(+8)**						
	10%	**dG(+7)_N3**						
	35%	**dT(+6)_O2**						
	6%	**dC3'(+8)**	✓	**A 424**	--			
	3%	**dC3'(+8)**	✓	**N 425**	--			
	4%	**dC(+7)**						
	7%	**dG(+7)_N2**						
			✓	**C 426**	--			
			✓	**L 430**	✓			
			✓	**I 431**	✓			
	90%	**dT(+5)**	✓	**V 432**	✓	**dT(-5)**	77%	
WAT			✓	**T 433**	✓			3x WAT
			✓	**I 464**	✓			
	100%	**dT(+4)**	✓	**S 465**	✓	**dT(-4)**	100%	
	93%	**dT(+3)_O4**	✓	**N 466**	✓	**dT(-2)_O4**	95%	
	65%	**dG(+2)_N7**				**dT(-3)_O4**	90%	
	44%	**dG(+2)_O6**						
			✓	**I 467**	✓			WAT-bridge
			✓	**C 468**	✓			
WAT-bridge	92%	**dT(+4)**	✓	**Q 469**	✓	**dT(-4)**	95%	
			--	**N 553**	✓			
			✓	**D 570**	✓			2x WAT
			--	**K 573**	✓	**dG(-1)**	2%	
			✓	**K 574**	✓	**dG(-1)**	84%	
			--	**E 616**	✓			3x WAT
			--	**K 642**	✓	**dT(-2)**	78%	
			✓	**Q 643**	✓			

In terms of the uSTAT3β-DNA complex, 23 corresponding residues were identified at the protein-DNA interface, with a further ten non-corresponding residues of monomer-A (G342, N420, G421, G422, A428, I464, D570, E616, Q643 and N646) and three non-corresponding residues of monomer-B (M329, R414 and N425). The observed asymmetry within the unphosphorylated monomer-DNA interaction is even more amplified here, indicating that monomer-A inserts deeper into the groove, which is in accord with the structural stability behavior described above (RMSF plots Figure 4.5 and 4.9) Overall, there is agreement for 34 protein-DNA interfacial residues between the two STAT3 complexes, with five extra residues being present at the phosphorylated complex interface, namely G419, A424, C426, N553 and K573; and two additional residues within the interface of at the uSTAT3-DNA complex, A428 and N646.

When analyzing the interfacial contacts, particular attention was paid to mapping the hydrogen bonds formed between the sugar-phosphate backbone of the DNA and protein residues, and in particular to those hydrogen bonds that involve DNA bases and/or have been reported to be affected by point mutations.[175,192] (Figure 4.11). Since the hydration of the protein surface may well be crucial for the stability and flexibility of the protein itself, the folding process, and molecular recognition, both water-protein and water-DNA interactions were analyzed, as well as features of protein-DNA recognition. The analysis of the hydrogen bonds formed between the solvent and STAT3-DNA interface throughout the MD simulation (1920 frames with 25-ps time-steps) used the VMD program. The water molecules were analyzed in terms of overall time spent within a distance required for a hydrogen bond formation, and only those water molecules with residence time exceeding 1 ns, were considered. The hydration at the protein-DNA interface was also analyzed in terms of water density maps (Figure 4.12 a-f).

4.4.4.1 Hydrogen bonds at the pSTAT3 protein-DNA interface

Following these criteria, the STAT3-DNA-solvent hydrogen bonds were analyzed in both phosphorylated and unphosphorylated STAT3 complexes. The main focus here is on the pSTAT3-DNA complex interface (Figures 4.11 a, 4.12 and Tables 4.2 and 4.3), so

the results on it will be discussed primarily, while the complimentary uSTAT3-DNA interfacial residues will be outlined subsequently. Nine protein residues were identified as forming hydrogen bonds involving the sugar-phosphate atoms of the DNA backbone, or the atoms from the bases, at both monomer interfaces, namely K340 (G(0) and G(-1)), Q344 (bp±1, ±2), R382 (bp±4), R417 (bp±3, ±4), R423 (bp±5, ±8), V432 (bp±5), S465 (bp±4), N466 (bp±2, ±3) and Q469 (bp±4). Monomer-A further formed stable hydrogen bonds between residues R414 – T(+3), N420 – C3'(+8) and G(+7), G421-C3'(+8), G422 – C3'(+8), R423 – A(+6), T(+6) and G(+7), A424 -C3'(+8), N425 – G(+8), G(+7) and C(+7). Protein residues are always hydrogen bond-donors and the DNA bases are hydrogen bond-acceptors, with the exception of G(+7) and C(+8) that behaved both as hydrogen bond-donors and acceptors (Figure 4.13 a). Monomer-B exclusively formed hydrogen bonds with the DNA sugar-phosphate backbone or the atoms from the bases; M331 – C(0), H332 – C(0), K573 – G(-1), K574 – G(-1), and K642 – T(-2). Despite fewer hydrogen bonds being formed between monomer-B and the DNA, the occupancy period of these interactions was in general longer comparing to the numerous interactions formed by monomer-A. This is due to the greater mobility of the DNA-binding domain of monomer A, allowing the hydrogen bond-interacting residues in loops ab, cx, ef and gα5 to reach into the major groove while the interface with the monomer-B is less variable as the simulation time progresses (Figure 4.11). R382, V432, S465, Q469 and N466 in both monomers formed hydrogen bonds with the DNA that lasted practically throughout the entire simulation time, with N466 specifically interacting with both DNA chains. These five hydrogen bonds are in accord with the experimental data, since they are observed at the x-ray structure of the STAT3-DNA complex (PDB id 1BG1), however another newly described interactions arise from two arginine residues R417 and R423; R417 of both monomers-A and -B forming long-lasting hydrogen bonds with T(±)3 and T(±)4, while R423 of monomer-A, specifically recognizing bases T(+6) and G(+7) at the complementary DNA strand, and further forming hydrogen bond with the terminal base at the C3' end of the opposite DNA chain, and A(+6) and T(+5), acting both as a hydrogen bond donor and acceptor; Also R423 of monomer-B formed hydrogen bonds with both chains of the DNA, namely T(-5) and T(-8); N425 of monomer-A specifically recognized base G(+7) with the DNA base behaving like hydrogen bond donor.

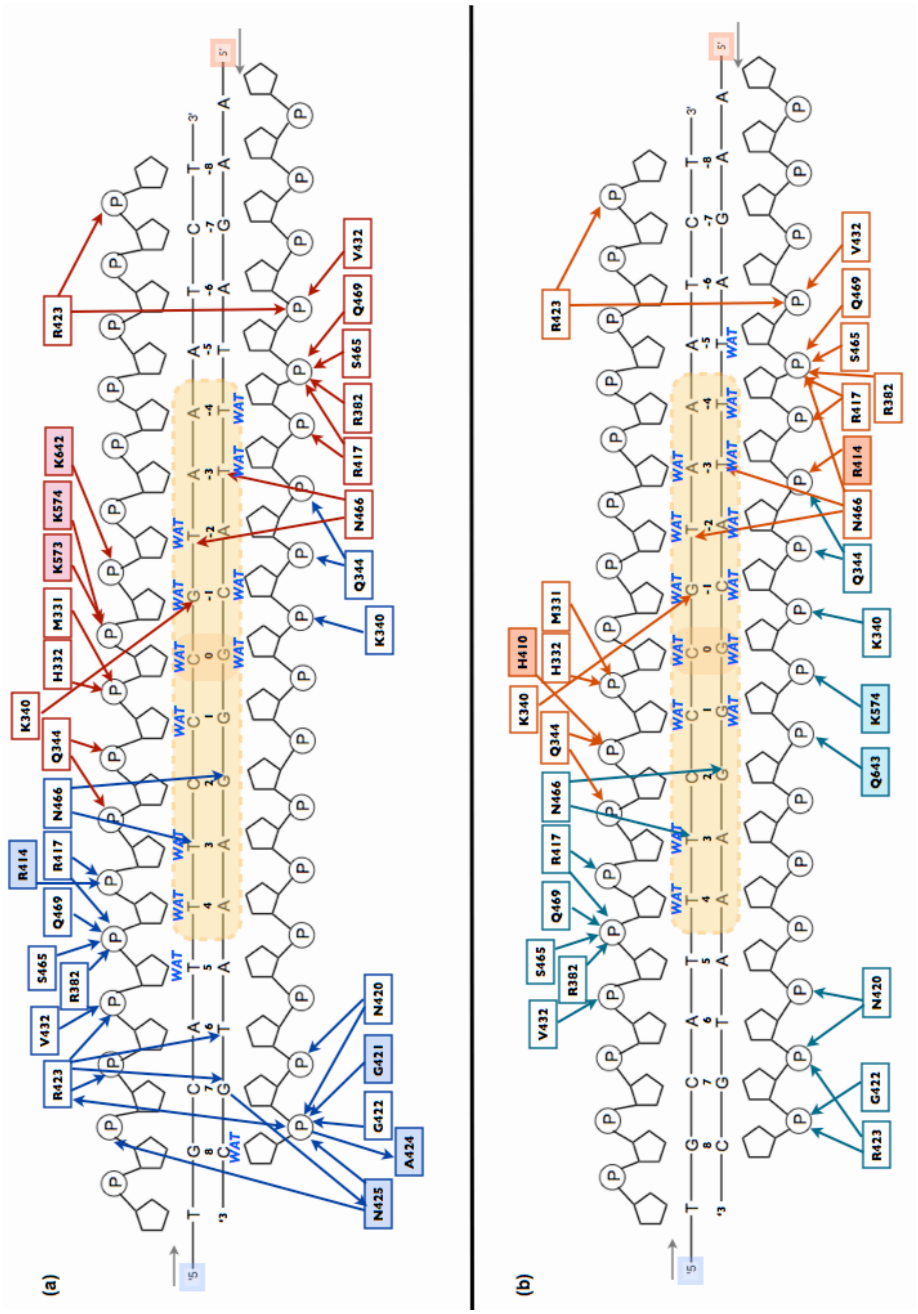


Figure 4.11: STAT3-DNA-water interaction maps formed via hydrogen bonds.
for the (a) pSTAT3-DNA complex and (b) uSTAT3-DNA complex; Residues of both STAT3 monomer-A (*blue/cyan*) and monomer-B (*red/orange*) forming direct hydrogen bonds with the individual nucleotides of the DNA duplex over the course of the MD simulation are represented. The 9-bp high affinity site is highlighted as a yellow bar. Arrows pointing toward/from the DNA backbone (represented by the phosphate group in the diagram) and the DNA bases represent the hydrogen bond formed. Water molecules identified to form hydrogen bonds with the DNA with overall residency longer than 1 ns are shown (labelled as WAT). Protein residues with colour-filled fields are those forming hydrogen bonds unique for each STAT3 complex.

The most frequently occurring residues in the protein-DNA interface are Arg, Asn, Lys and Gln residues, with Arg residues located in the AT-rich minor groove, which is in accord with findings of arginine-enriched DNA minor grooves reported by Rhos[211] *et al*. There is also a large number of Gly residues at the monomer-A-DNA interface, in particular forming hydrogen bonds with the terminal base at the C3' end of the DNA strand.

There are a total of 25 contacts formed between 21 different water molecules and 12 individual nucleotides (+5, +4, +3, +1, 0, -1, and -2 in the 5'-T strand 1; -4, -3, -1, 0, and +8 in the 5'-A strand 2). 16 of these 21 water molecules (water molecules are numbered 1-16 for simplicity - summarized in Table 4.3) bridge to the interfacial protein residues, forming indirect hydrogen bonds, for instance with E415 and T341 in both monomers, and additionally with Q344 and Q469 of monomer-A, and K340, V343, and I467 in monomer B. There is a pronounced pattern in the frequency of interactions of base pairs T±4, T±3 with both protein residues and water molecules, and there are also numerous hydrogen bonds involved with bp0 and bp±1, despite the high sequence variability at these positions in natural DNA target sites.

Table 4.3: Bridging water molecules at the protein-DNA interface of the pSTAT3-DNA complex.
These bridging water molecules mediate the indirect hydrogen bonds formed between the residues of monomer-A (*white*) and monomer-B (*grey*) and the DNA

pSTAT3-DNA complex-bound monomer	Bridging waters	pSTAT3-DNA complex-bound *ds*DNA
E415	WAT 1	**dT(+4)**
E415, R282, T341	WAT 2	**dT(+4), dT(+3)**
E415	WAT 3	**dT(+4)**
Q344	WAT 4	**dC(-1)**
E415	WAT 5	**dT(+4)**
Q469	WAT 6	**dT(+5)**
Q344	WAT 7	**dC(-1)**
K340, I467	WAT 8	**dC(0), dG(-1), dT(-2)**
E415, T341	WAT 9	**dT(-3), dT(-4)**
T341	WAT 10	**dT(-3)**
K340, V338, E415	WAT 11	**dC(0)**
E415	WAT 12	**dT(-4)**
E415	WAT 13	**dT(-4)**
E415	WAT 14	**dT(-4)**
E415	WAT 15	**dT(-4)**
K340, V343, L413	WAT 16	**dC(0)**

4.4.4.2 Hydrogen bonds at the uSTAT3 protein-DNA interface

Among the 36 protein residues identified at the protein-DNA interface, 34 of which are corresponding to the interfacial residues of the phosphorylated complex, the same nine protein residues, as in the case of the pSTAT3 complex, were identified forming hydrogen bonds involving the sugar-phosphate atoms of the DNA backbone, or the atoms from the bases, at both monomer interfaces (i.e K340, Q344, R382, R417, R423, V432, S465, N466 and Q469). In monomer-B, an additional hydrogen bond between residues N466 and T(-4) was also identified. (Figure 4.11 b). Monomer-A further formed hydrogen bonds between residues N420 - T(+6) and G(+7), G422 - C3'(+8), K574 - G(+1) and Q643 - G(+2), where protein residues are always hydrogen bond-donors and the DNA bases are hydrogen bond-acceptors. In comparison to the phosphorylated monomer-A interactions, hydrogen bonds involving residues R414, G421, A424 and N425 are not present here. Monomer-B exclusively formed hydrogen bonds with the DNA sugar-phosphate backbone or the atoms from these bases: M331 - (0), H332 - C(0), H410 - C(+1) and R414 - A(-2); the first two of which are corresponding to those observed in phosphorylated monomer-B, while the latter two are unique. Hydrogen bonds comprising residues K573, K574 and K642 were not present here in the frame of selected criteria. In this case, an equal number of hydrogen bonds formed between the unphosphorylated complex-forming monomers and the DNA, in particular 18 hydrogen bonds arising from 13 residues of each of the monomers (Figure 4.11 a and b for comparison). This is in accord with the structural stability data and the PCA analysis (Figure 4.9 d-f), which revealed overall very similar scope of movement among the DNA-binding domains, monomer-A being dominant in terms of insertion into the groove of the DNA at its C3' terminal, hence forming contacts that are not observed at the T3' end of the complimentary DNA chain. By comparison with the pSTAT3 complex, there are a total of 25 contacts formed between 19 different water molecules and 14 individual nucleotides (+4, +3, +1, 0, -1, -2 and -3 in the 5'-T strand 1; -5, -4, -3, -2, -1, 0, and +1 in the 5'-A strand 2). 15 of these 19 water molecules bridge to the interfacial protein residues, forming indirect hydrogen bonds, for instance with K340, T341, Q344, R382, E415, I467 and Q469 in both monomers, and additionally with M331 and H332 of monomer-A.

This study shows that the solvent molecules are indeed crucial for specific STAT3β-DNA recognition, expanding the contact area formed between the protein and duplex DNA. For comparison, the same water-DNA contact analysis performed with the 17-bp DNA alone suggest that only two water molecules interact with A(+5) and T(+4) of the opposite DNA strand, and do not exceed 2% of the overall occupancy time spent within that hydrogen bond. In terms of the U-STAT3 monomer, only one (E616) out of many residues previously determined to be at the interface of the STAT3-DNA complex in interaction with a solvent molecule, was found to actually form a corresponding hydrogen bond with a water molecule, again with a shorter occupancy time. Three more interfacial residues determined from the STAT3-DNA complex were found to form hydrogen bonds with water molecules in U-STAT3, namely A424, C426, F384 (with occupancy times not exceeding 3% of the analyzed 48 ns). Hydration maps representing the water density at bp±4 are shown in Figure 4.12 (c, d) (pSTAT3 complex only), where numerous solvent molecules forming hydrogen bonds with the nucleotide were identified. (as also shown in Table 4.3). Three hydrogen bonds were also found to be formed between C(0) and the solvent molecules (Table 4.3 and Figure 4.12 e), or between K340 of monomer B with the solvent molecules (Figure 4.12 b).

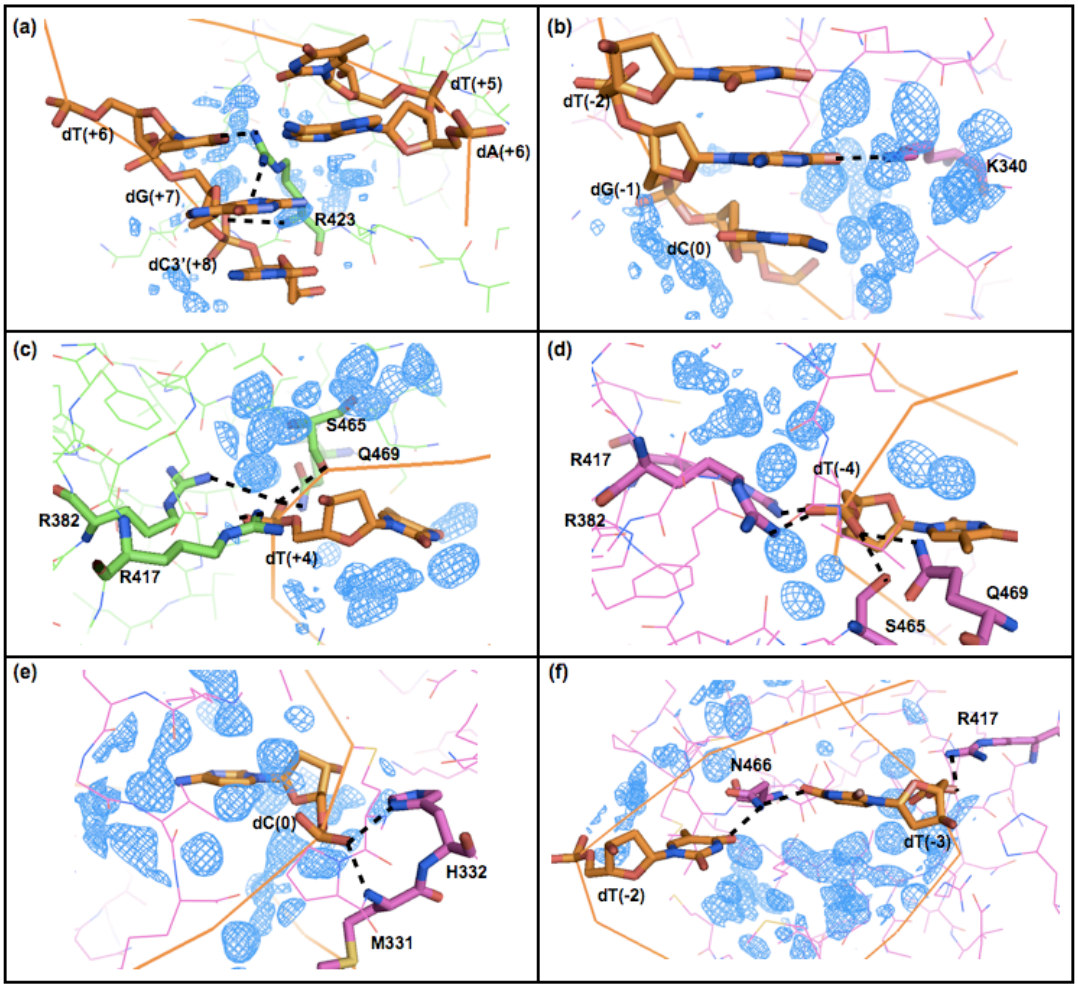


Figure 4.12: Water density maps showing the hydration at the pSTAT3 protein-DNA interface. Protein residues of monomer-A are displayed in *green*, monomer-B in *magenta,* and the *ds*DNA bases in *orange.* Regions with high water density, as observed throughout the MD run, are shown in *blue* "blobs" in mesh representation, where the size of the blobs corresponds to sizes of densely hydrated regions. Densely hydrated regions were defined as regions commonly occupied by water molecules throughout the simulation time, but they were not determining the exact number of water molecules present at the region (blob).

4.4.5 Locating the mutations in the protein-DNA interface

31 different dominant-negative STAT3 mutations have been identified to date,[175,192,191] affecting predominantly the DNA-binding and SH2 domains of STAT3. The clinical importance of these mutations, accounting for the vast majority of Hyper-IgE Syndrome (HIES) cases, has been reported and summarized.[176] Since the focus of this study is on the protein-DNA recognition aspects of STAT3, those point mutations affecting the DNA-binding region (17 mutations) were examined for possible structural explanations of these mutations, which are summarized in Table 4.4. There is a level of symmetry between the monomers in terms of the residues affected by the point mutations. Thus R382 has the largest frequency of mutations observed in HIES patients, and it affects hydrogen-bonded DNA recognition with V463 and E435. Four different mutations of N466 have been reported, with low frequency of occurrence but with crucial effects on the direct readout involving bases at both strands of the DNA, and further interactions with C468 and Q469 (~ 65% and ~ 35% HB occupancy respectively), causing a cascade of the subsequent alterations in STAT3-DNA recognition. RMS fluctuations on a per residue basis, for the DNA-binding region, have been examined as a comparison between the phosphorylated complex-bound monomers-A and -B and unphosphorylated U-STAT3 monomer, to explore the dynamics of the interfacial residues, and in particular the residues for which mutations were observed. However there is not a significant correlation between the residues with larger values of RMS and the occupancy time they spent forming hydrogen bonds with the DNA (Figure 4.6), and so these are not discussed further. In order to better map out the positions of the interfacial residues that may be mutated in terms of the phosphorylated complex-bound STAT3 and U-STAT3 monomer, the first eigenvectors for the backbone atoms were structurally aligned (duplex DNA was kept in place as a reference). The locations of the affected residues are in good agreement in the two structures, with the exception of R423 which is further away from the groove in the case of the U-STAT3 monomer (Figure 4.13). This may be explained by the high affinity of the residues within the ef-loop of the complex-bound STAT3 for DNA, hence having a large effect on the mobility of that particular region. Overall, the most significant structural difference between the two monomers is at the protein-protein interaction region, in agreement with the results discussed above.

Table 4.4: Point mutations within the DNA-binding region and their interactions*.

Point mutation	Patients reported	HB affected in monomer-A	HB affected in monomer-B
H332Y	3	**M329, D334, R335, WAT**	**dC(0), M329, R335, K573**
R335W	2	**H332, D334, D566, 2x WAT**	**M329, H332, D334, D566, S574, 3x WAT**
K340N	1	**dG(0), M329, V343**	**T341, V343, WAT8, WAT11, WAT16**
G342D	1	**L413**	**L413**
V343L	1	**K340, T412**	**K340, T412, WAT16**
R382L	2	**dT(+4), E435, V463, WAT2, WAT**	**dT(-4), E435, V463**
R382W	35		
R382Q	14		
F384L	2	**2x WAT**	**D369, K383**
F384S	1		
T389I	1	**K409, H410**	**---**
T412S	1	**L387, V343**	**V343, L387, L411**
R423Q	6	**dA(+6), dT(+6), dT(+5) dG(+7), dC3'(+8), G380, WAT**	**dT(-5), dA(-6), dT3'(-8) G380, A428, L430, E435**
V432M	1	**dT(+5), E435**	**dT(-5) , E435**
H437P	1	**D369, E435, V461**	**D369, E435, V461**
H437Y	1		
S465A	2	**dT(+4), Q469**	**dT(-4), Q469**
N466D	1	**dG(+2), dT(+3), C468, Q469**	**dT(-2), dT(-3), C468, Q469**
N466S	1		
N466T	1		
N466K	1		
N466H	1		
Q469H	1	**dT(+4), S465, N466, A473, WAT6**	**dT(-4), S465, N466, A473**
N472D	1	**S476, 2x WAT**	**S476, 2x WAT**
K642E	1	**N567, D570, E616, L645, N646**	**dT(-2), D570, E616, L645, N646**

*All hydrogen bonds formed between the mutated residue and any other protein residue, DNA base, or solvent molecule throughout the course of the simulation, are listed. Bridging water molecules involved in the interactions are specifically labeled (corresponding to those in Table 4.3).

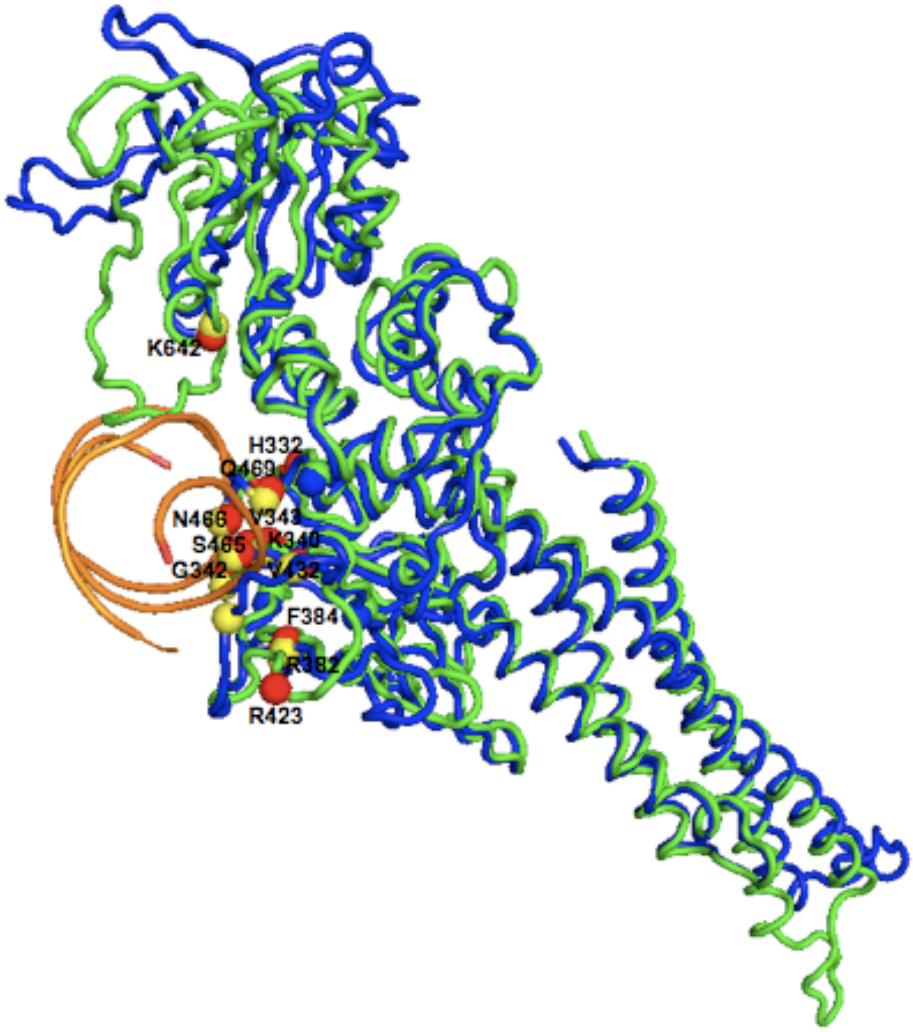


Figure 4.13: Structural alignment of the first eigenvectors with locations of point mutations.
For the backbone atoms of the pSTAT3 monomer-A (*blue*) and unphosphorylated STAT3 monomer (*green*) with the single point mutations at the DNA-binding region highlighted in sphere representation (*yellow* for monomer-A, *red* for U-monomer). The DNA duplex (*orange*) corresponding to the conformation of the first eigenvector (for the STAT3 complex) is displayed in order to better define the orientation of this view.

4.5 CONCLUSIONS

(I) Despite their sequence and secondary structure identity, distinct dynamic properties/ behavior was observed for monomers A and B within both phosphorylated and unphosphorylated STAT3β-DNA complexes, and their characteristic dynamic properties contrast with that of the unphosphorylated STAT3 monomer. There are differences in their magnitude of motion, among both complexes and with respect to the unbound latent STAT3 monomer, in particular in the SH2 domain with the stretch of C-terminal residues, that clearly folds in the case of U-STAT3.

(II) Examination of protein-DNA-solvent interactions focussed on those water molecules that remained at the interface with an occupancy greater than 1 ns of the simulation time. Water molecules were shown to be important mediators of protein-DNA recognition; there were a number directly involved in recognition, forming stable multiple hydrogen bonds with the protein and DNA residues. Bridging water molecules were identified not only in the protein-DNA interfacial interactions (residues within 5.0 Å of the DNA), but also for instance between V338(B)- C(0), and L413(B) – C(0). While the water-DNA interactions in both phosphorylated and unphosphorylated STAT3 complexes were stable over the simulation time, only two water molecules forming hydrogen bonds with the duplex DNA alone were identified which did exceed the overall 1-ns occupancy (hydrogen bonds with A(+5) and T(+4)). The simulation of the duplex DNA alone did not show unwinding at the terminal regions on the time scale of the simulations, despite the 5' overhanging ends. The protein dimer does then have a stabilizing effect on the DNA, although a small amount of DNA unwinding was observed at the complex-bound DNA, at the T5’ region as a consequence of the multiple interactions of monomer A with the opposite chain of the DNA in the region C3'(+8), G(+7). In case of the unphosphorylated STAT3 complex, the protein-DNA interaction described by means of hydrogen bond formation is not so strong at the C3’ region, spanning less residues and DNA bases than in the phosphorylated complex.

(III) Experimentally determined point mutations[191,192] in the DNA-binding domain were located at the protein-DNA interface.

(IV) Molecular dynamics provides more detailed insight into the structural description of the protein-DNA recognition compared to the only currently available crystal structure of the STAT3β-DNA complex,[181] which has been used to date for docking studies of STAT3 SH2 small-molecule ligands. These new data should therefore provide a more-robust platform for *in silico* approaches to the design of STAT3-STAT3 inhibitors.

CHAPTER 5:
Characterization of molecular recognition of STAT3 SH2 domain, and its small-molecule inhibitors by means of combined *in silico* approaches

5.1 BACKGROUND

There has been considerable progress in recent years in terms of the disruption of dimeric transcription factors (such as STAT3) with small molecules that directly target the protein-protein interface of the protein complex. However, the modulation of protein-protein interactions (PPIs) for therapeutic intervention poses significant challenges, with respect to the structural features of proteins and the chemistry of their interactions.[212] Among the identified general PPI targeting hurdles are: extensive contact surfaces (1,500-3,000 Å^2); lack of suitable pockets for small-molecule ligands despite the interfaces being studded with clefts and indentations; and the nature of their predominantly hydrophobic interactions. Thus the physico-chemical properties of the potential small-molecule inhibitors that would maximize binding complementarity appear to be demanding and non-drug-like (high molecular weight, large hydrpohobic surfaces). Furthermore, PPIs involving Src Homology 2 (SH2) domains, such as STAT3, can be particularly challenging for pharmacological intervention as they are characterized by highly-polar phosphotyrosine (pY) residues.[181]

5.1.1 Targeting STAT3 SH2 domains for therapeutic intervention

SH2 domains are ubiquitously present within signal transduction proteins (such as STATs) and they mediate protein-protein interactions (PPI) by recognizing specific phosphotyrosine (pY) sequences on a target protein.[213] SH2 domains represent the largest class of pY-selective recognition domains in the human proteome, and they are

highly conserved among all family members of tyrosine kinases.[214] It is known that compounds targeting the STAT3 SH2 domain may inhibit STAT3 function by (1) preventing docking to cell surface receptors, thus impeding phosphorylation of Y705, subsequent dimerization, nuclear translocation, and gene expression; and by (2) disrupting STAT3 dimers thus preventing translocation to the nucleus and DNA binding, leading to inhibition of downstream gene expression involved in survival, cell cycling, or angiogenesis.[215]

Whereas most groups have developed inhibitors that block the SH2 domain, the DNA-binding domain and N-terminal domain have also been targeted to inhibit DNA binding and nuclear translocation respectively. Because of this, a very diverse range of molecular inhibitors have been patented for STAT3 protein function inhibition.[216] An overview of these inhibitors, ranging from peptidomimetics and oligonucleotides to small molecules and platinum-based compounds is given in CHAPTER 1. Here I will only provide an outline of computational studies, where molecular docking has been employed to rationalize and rank known biologically active STAT3 ligands (in the absence of structural data), and has provided further insight into their complexes with STAT3 SH2 domain by means of MD simulations and subsequent binding energies calculations.

5.1.2 Dynamic aspects of STAT3 studies: experimental and *in silico* view

A dynamic view of a macromolecular target for drug discovery is essential in order to obtain valuable insight into its behavior. The dynamic aspects of STAT3 behavior have been studied experimentally, at the STAT3 signaling pathway level, as well as computationally, via molecular dynamics simulations, at the level of a single STAT3 macromolecule or its complex. For instance, Watanabe *et al*[217] investigated the mobility and dynamics of STAT3 in IL-6 signaling in living cells by means of fluorescence correlation spectroscopy, employing a STAT3-GFP hybrid. Mohr *et al*[218] revised all the aspects of dynamics and non-canonical JAK/STAT signaling, supporting the observations of unphosphorylated STAT3 not only forming a homodimer, but also translocating to the

nucleus and binding specific DNA sequences.[179] Now that the original STAT3 signaling dogma has been challenged by these observations, it follows that both phosphorylated and unphosphorylated STAT3 dimers should be taken in account for small-molecule inhibitors design. Molecular docking studies of phosphorylated and unphosphorylated STAT3 SH2 domain forms will be presented and discussed in the following sections.

In terms of *in silico* dynamics studies, comparative explicit solvent molecular dynamics study of STAT3 and STAT1 homodimers has been carried out by Lin *et al*[183] with the focus on conformational changes of the two studied complexes over a 50-ns frame at 310 K, employing NAMD[219] with CHARM27[159] force field. In contrast, a comparative explicit solvent MD study of STAT3-DNA complex, with respect to its latent unphosphorylated monomer, has been reported by Husby *et al*[220] using the GROMACS program and employing the AMBER force field. Comprehensive molecular docking studies, employing AUTODOCK[221] v 4.2 and VINA v 1.1, were carried out by Dhanik *et al*[222] to identify the most potent STAT3 SH2 domain ligand out of the 142 currently known peptidomimetic inhibitors. The authors reported ~65% correlation between predicted binding energies and experimental IC_{50} values, using a single conformation of the STAT3 SH2 domain obtained from the STAT3 X-ray structure.[181]

5.1.3 Origin of the "ESP" library of small molecules for molecular docking study

A medium throughput screening of ~ 25.000 biologically active and chemically diverse small molecules (i.e Diversity Set library), provided by the Evotec drug discovery company, was performed at the European Screening Port (ESP) in Hamburg, Germany. Initially, a biochemical cell-free Fluorescent-Polarization (FP)-based primary PPI binding assay was carried out at a fixed 40 μM compound concentration, which formed 223 hits that showed > 50% inhibition. Out of the 223 hits, only 54 were commercially available compounds and purchased by the CRUK PPI Drug Discovery research group at the UCL School of Pharmacy. A dose response FP analysis was carried out on those 54 compounds, leading to a selection of six compounds (Figure 5.1), based on their optimal dose response curve shapes (MTS assay-based). The next step, currently still in

process, is the assessment of STAT3 transcriptional activity inhibition in cells via a STAT3 luciferase reporter assay, with the SV40 luciferase assay being employed as a control (Supplementary Figure S5.1).

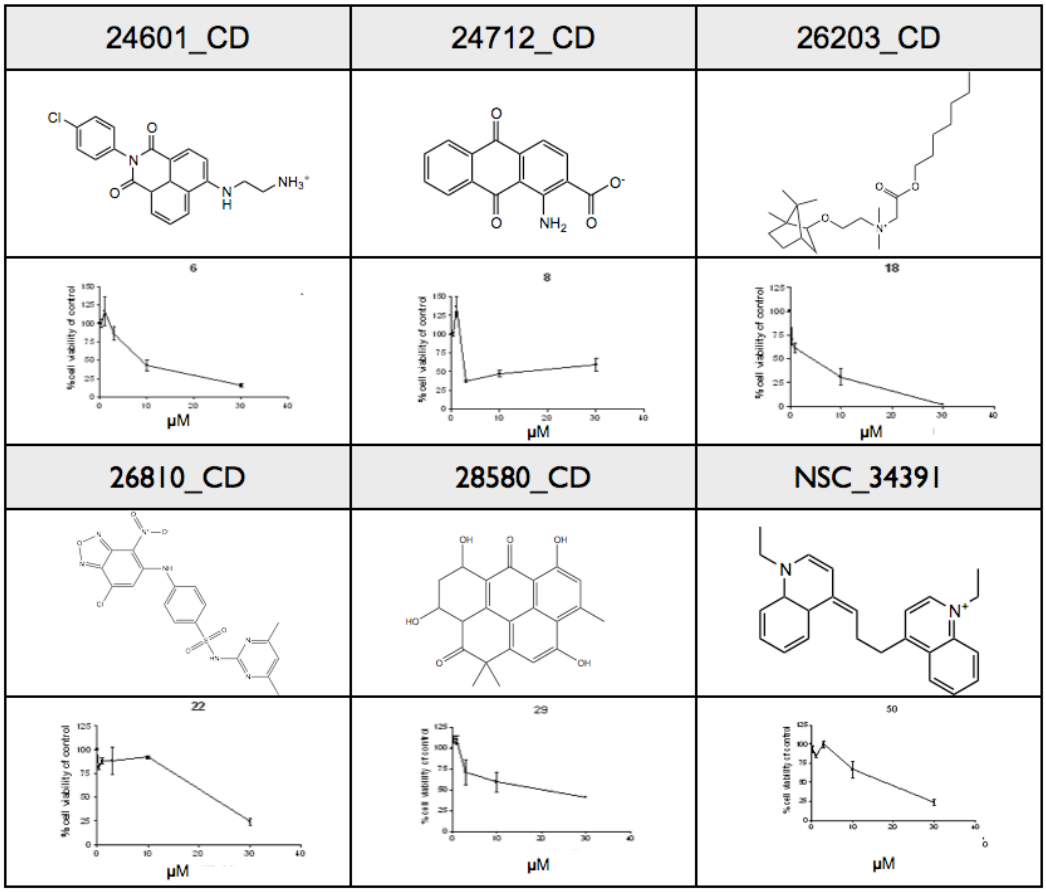


Figure 5.1: Six compounds selected from the ESP HTS with optimal MTS dose-response curve shapes. Chemical structures together with the MTS-assay graphs are shown for each of the experimentally selected molecules.

The 54 compounds were also used within a comparative STAT3 molecular docking study, in order to gain insight into their possible modes of interaction with the STAT3 SH2 domain. Recent studies revealed that state of the art docking algorithms predict an incorrect binding pose for about 50-70 % of ligands when only a single fixed receptor conformation is considered.[141] However in the present study, multiple conformations of the receptor, SH2 domain of both phosphorylated and unphosphorylated STAT3, were used to partially overcome the numerous computational docking challenges, and to obtain more accurate docking predictions.

5.2 AIMS

Work was carried out simultaneously with the ongoing MD simulations of the STAT3βtc systems as described in CHAPTER 4, so the MD simulations setup and explicit solvent trajectory calculations will not be described here. However, their exact trajectories (of the pSTAT3 and uSTAT3 complexes), previously analyzed in terms of protein-DNA interactions, were employed here with the focus on the PPIs mediated by the SH2 domains. The ultimate goal is to study and understand the shape and difference in their respective pY705/Y705 binding pockets. This in turn provides a platform for a comparative molecular docking study from various perspectives; i.e. two molecular targets whose only difference is phosphorylation of the incriminated tyrosine residue (Y705); use of multiple representative structures (i.e multiple receptor conformations, MRC); and a comparison of two well-known and widely-used molecular docking software packages. Insight into the chemical features that are specific for the pY recognition by opposite SH2 domain, as well as into thermodynamic stability of the STAT3:STAT3 complexes, respective to their phosphorylated- and unphosphorylated-form, then completes the "picture". The main questions set out to be answered within this combined-*in silico* approach are:

• What are the structural differences in the shape of the pY705 and Y705 binding pockets at the protein-protein interface of the pSTAT3 and uSTAT3 complexes respectively?

• What are the specific chemical features of the pY705-containing tetrapeptide interactions with the reciprocally-bound SH2 domain, and how stable they are?

• How can simultaneous use of two well established molecular docking software packages, combined with a set of multiple receptor conformations enhance a successful prediction for a potential PPI inhibitor?

• What is the binding affinity of the pSTAT3 protein-protein association with respect to uSTAT3, and what is the difference in their respective free binding energy at the protein-DNA level?

5.3 METHODS

Two parts of the explicit solvent MD trajectories of the pSTAT3 and uSTAT3 complex were employed to collect multiple representative structures (multiple receptor conformations, MRC), that were then employed within the subsequent studies:

• 25-ns MD-trajectories of the pSTAT3 complex and uSTAT3 complex were used for collection of representative MRC (SH2 domains) for a subsequent comparative molecular docking study with over ~50 biologically active compounds, employing two established, constantly improving, and popular molecular docking suite of programs: GOLD[223] and DOCK6[224](section 5.3.1 - 5.3.3)

• full-length 50-ns MD-trajectories (48-ns respectively) were utilized for the ensuing energetic analysis of the protein-protein, and protein-DNA interaction in both pSTAT3 and uSTAT3 complex (section 5.3.4); and for the 3D-pharmacophore modeling of phospho-Tyr705 (pY705) interactions in the pSTAT3-DNA complex (section 5.3.5);

5.3.1 Preparation of the multiple-target conformation via cluster analysis

Trajectories of the pSTAT3 and uSTAT3 complex, sampled over 25-ns and 50-ns timeframes respectively, were examined by a clustering agglomerative algorithm, in order to collect multiple target conformations for molecular docking and 3D-pharmacophore modeling study, on the basis of the following scenario:

(1) 25-ns trajectories of the pSTAT3 complex, and uSTAT3 complex, were clustered with a RMSD cutoff distance of 2.5 Å. The entire STAT3-DNA complex (e.i residues 136-716 and dsDNA) was considered for clustering, and the three most populated clusters of the pSTAT3 and uSTAT3 complexes, represented by a middle structure sampled at 300K, were chosen for comparative molecular docking study.

(2) the 50-ns (48-ns respectively) trajectory of the pSTAT3-DNA complex was clustered applying the RMSD cut-off distance of 2.0 Å for two structures to be regarded as neighbors. Only the SH2 domains with the stretch of C-terminal residues (i.e residues 586 - 716) were considered for clustering here, to follow up on the protocol described in previous chapter (section 4.3.3).

5.3.2 Ligand preparation

Minimized structures of 54 biologically active compounds, - an 'ESP-library' selected by a medium throughput screening completed at the European ScreeningPort laboratories in Hamburg (and purchased for further biological evaluation carried in our laboratories) were constructed with ChemBioOffice (www.cambridgesoft.com), exported in pdb format, and subsequently converted into the Sybyl molecule mol2 file format, to be used as input for the GOLD[223] and DOCK6[224] molecular docking simulations. The ligand structures were protonated, assigned atomic partial charges employing the *AM1-BCC*[225] charge (i.e atomic charges with simple additive bond charge corrections) calculation method within Chimera's *Dock Prep* module, and their atom types (AMBER GAFF[226]) were assigned using the ANTECHAMBER[170] program as implemented in UCSF Chimera.[208]

5.3.3 DOCK6 docking protocol

The DOCK algorithm,[227] implemented into the DOCK6 suite of programs, addresses rigid body docking using a geometric matching algorithm to superimpose the ligand onto a negative image of the binding pocket of the macromolecular receptor. An algorithm for flexible-ligand docking, on-the-fly optimization and improved algorithm's ability in finding the lowest-energy binding mode of the small-molecule ligand, together with free academic licensing makes DOCK6 a popular choice for molecular docking (and virtual screening) studies. DOCK6 was the primary choice for the docking study with multiple target conformations of the SH2 domain, sampled through MD simulations of the pSTAT3 and uSTAT3 complexes. The simulation protocols were kept identical for the consistency of the results and their subsequent comparison.

Several steps of receptor preparation were necessary prior to docking of the 54 small molecules with the three conformations of the pSTAT3βtc-SH2 domain and three conformations of the uSTAT3βtc-SH2 domain (residues 586-688) in DOCK v 6.4;[224]

• Receptor structures were processed with the *Dock Prep* module of Chimera, through which hydrogens at the terminal residues were added, and protein residues were assigned AMBER ff99SB partial charges. The receptors were then output in mol2 and pdb file format.

• A molecular surface for the receptor was generated by the *dms* program (based on the algorithm developed by Richards and adapted by Connolly[228] by rolling a ball the size of a water molecule over the van der Waal's surface of the receptor. Simultaneously, the surface normal vector at each surface point was computed for the subsequent sphere calculations.

• The binding site was represented by a set of spheres (with the minimum and maximum sphere radius of 1.4 Å and 4.0 Å respectively), selected within 10.0 Å from the structure of the 'natural ligand', represented by a pentapeptide P704-(p)Y705-L706-K707-T708 of the opposite STAT3 monomer (Figure 5.2). The pentapeptide was prepared following the procedure described above, - hydrogens were added and standard amino acids were assigned AMBER *ff99SB*[167] partial charges, while *AM1-BCC*[225] charges were computed for the non-standard pY residue. A binding site of each of the six receptor conformations was in an average defined by 68 ± 8 spheres. Small-molecule ligands were then automatically orientated into those spheres, cycling through a maximum of 500 orientations (default value is 100), to allow a generous conformational search.

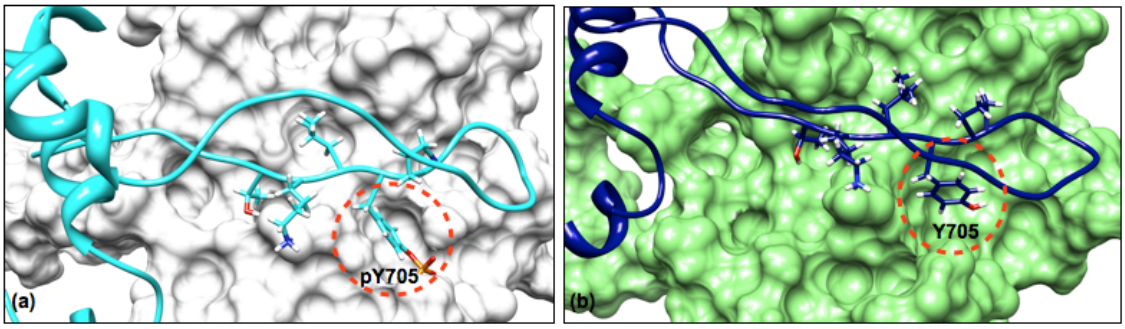


Figure 5.2: Defining the binding site of the PPI, via the SH2 domains of the STAT3βtc complexes. (a) pY705-containing loop of an SH2 domain (pSTAT3), shown in ribbon representation (*cyan*), binds to the opposite SH2 domain, shown in molecular surface representation (*white*); (b) Y705-containing loop of an SH2 domain (uSTAT3), shown in ribbon representation (*dark blue*), binds to the opposite SH2 domain, shown in molecular surface representation (*light green*).

• To account for the receptor contribution to the score (i.e force field based score) throughout docking, grids that store the van der Waals (VdW) and electrostatic components for the receptor were computed, allowing for a rapid energy score evaluation. The scoring grids were computed with 0.3 Å spacing between grid points along each axis, resulting on average to ~ 130 x ~130 x ~150 grid points per side (x, y, z).

• An *anchor-and-Grow*[224] algorithm for incremental ligand construction was employed to sample the internal degrees of freedom of each small molecule ligand (i.e. flexible ligand treatment). Upon the ligand's largest rigid substructure identification (defined by a minimum of five heavy atoms), and its rigid orientation in the binding site, the resulting orientations were ranked according to their score, spatially clustered and pruned (based on RMS-distance between each candidate and a top-ranked reference configuration). Subsequently the remaining flexible portions of the ligand were built onto the best anchor orientations and re-optimized.

• To discriminate among orientations of the molecules DOCK6 implemented scoring functions were employed. Prior to scoring, the orientations of the ligand which significantly overlapped receptor atoms were discarded (via the *bump filter*). Subsequently, grid-based primary energy scoring function was used during rigid orienting, anchor-and-grow steps, and minimization. For the final minimization, scoring, and ranking of the molecules, the Hawkins GB/SA[131,229] (GB model supplemented with the solvent accessible surface area SA term) secondary scoring function with a 150mM salt concentration was used (as an implementation of the MM-GBSA method) using the pairwise GB solvation model by Hawkins.[230,231] The total interaction between the ligand and receptor were then represented by unscaled Coulombic and Lennard-Jones 12-6 potentials (MM) with the change in solvation (GBSA). A total of three secondary scored ligand conformations were saved for each of the small molecule ligands, with the best-ranked conformation being considered for further analysis by means of Chimera v1.5 software.

To test and validate the choice of the docking protocol parameters, a phosphotyrosine residue was docked with the three target conformations of the pSTAT3 SH2 domain, and subsequently a tyrosine residue was docked with the uSTAT3 SH2 domain representations, into their respective binding sites defined by the PpY/YLKT pentapeptide. The top-ranked ligand poses were then visually assessed and compared.

5.3.4 GOLD docking protocol

GOLD, an automated ligand docking program that employs a genetic algorithm to examine the complete range of small-molecule ligand conformational flexibility in partially-flexible protein binding site. GOLD[223] (Genetic Optimisation for Ligand Docking), was the second program of choice for the comparative molecular docking study employing the multiple target conformation approach. The GOLD docking program has been in continuos development since it was first introduced more than 15 years ago, making GOLD one of the most widely used and cited docking programs currently available.[232,233]

Automated docking was carried out using GOLD v 5.0.1, with the binding site of each of the receptors' six total conformations (three pSTAT3βtc-SH2 domain, and three uSTAT3βtc-SH2 domain) defined as the residues with at least one heavy atom within 10 Å from a protein atom, NH1 of residue R609. The region of interest was restricted to concave, solvent-accessible surfaces, by applying the cavity detection algorithm, as implemented in GOLD. There was no water present at the binding site, and all missing hydrogen atoms (i.e at the terminal residues) were added. The default docking protocol was applied, employing the GoldScore fitness function, which has a van der Waals treatment of clash and dispersion terms and uses a crystal structure-derived treatment of hydrogen bonds (and also metal terms).[223] GoldScore comprises three main components: (1) a hydrogen bonding term (based on empirical values for the strength of hydrogen bonds between different atom types); (2) a van der Waals term (which accounts for for protein-ligand hydrophobic interactions) and (3) a ligand internal energy term (which comprises ligand internal vdw and torsional strain energy).[223] A genetic algorithm (GA), implemented within GOLD, was utilized to optimize the fitness score of the docked ligands. An evolutionary strategy was employed during exploration of the conformational variability of the ligand within the binding site. The most accurate (slow speed) parameter settings was chosen through the interactive GOLD interface, with the number of operations (i.e the number of operators that are applied over the course of the GA run) set to 100,000. Each docking was then performed ten-times on each ligand. The best pose, a predicted best ligand binding position assessed by a

dimensionless fitness score, was saved for a subsequent analysis. Only the torsions around the ligand's flexible bonds were optimized during docking (i.e flexible ligand treatment). In terms of the receptor, GOLD allows for partial protein flexibility, hence the torsion angles of residues Ser, Thr and Tyr hydroxyl groups (-OH) were allowed to rotate during docking, providing a better level of optimization of the hydrogen-bonding interactions of these residues with the ligand. A corresponding principle of optimization applied to NH_3^+ groups of Lys residues, otherwise all other parts of the receptor were kept rigid.

5.3.5 Intermolecular interaction energy calculations for the STAT3:STAT3 and STAT3dimer:DNA association

To obtain a quantitative estimate of the binding free energy of the pSTAT3βtc complex and uSTAT3βtc complex formation at both the protein-protein and protein-DNA level of molecular association, the interaction energy together with the solvation free energy were calculated employing the Molecular Mechanics/Poisson-Boltzmann (Generalized-Born) Surface Area (MM/PB(GB)SA) method implemented in AMBER 11.

MM/PB(GB)SA methods have been commonly used to investigate PPIs and other protein-ligand and/or protein-DNA interactions since they combine the speed of a continuum approach to modeling solvent interactions with the MM-based level of theoretical (and accurate) approach toward full-atomic modeling of the biomolecular interactions. The principle of this method can be well outlined by its abbreviation: MM stands for the molecular mechanics force fields employed to calculate both intermolecular and direct intramolecular contributions to binding free energies; PB and GB refer to the implicit solvent methods used to calculate the electrostatics contributions, and SA stands for solvent accessible surface area (SASA) methods used to calculate the non-polar contributions to binding free energies.[123] The entropic contributions are calculated separately, and may added in further refinement.

The interaction energies of the STAT3:STAT3 association (ΔG_{bind}(PPI)), and the STAT3dimer:DNA complex formation (ΔG_{bind}(PDI)), for both pSTAT3 complex and uSTAT3 complex, are described by calculating the Gibbs free energies for the complex, receptor and ligand individually, across the configurational ensemble obtained from the MD trajectory, according to the following equation:

$$\Delta G_{bind}(PPI) = G_{STAT3:STAT3} - (G_{STAT3\text{-}monoA} + G_{STAT3\text{-}monoB}) \quad (5.1)$$

$$\Delta G_{bind}(PDI) = G_{STAT3:DNA\ complex} - (G_{STAT3:STAT3} + G_{DNA}) \quad (5.2)$$

where $G_{STAT3:STAT3}$, $G_{STAT3\text{-}monoA}$, $G_{STAT3\text{-}monoB}$, $G_{STAT3:DNA\ complex}$ and G_{DNA} are the calculated average free energies of the STAT3βtc homodimer, STAT3βtc homodimer-forming monomers A andB, STAT3βtc homodimer:DNA complex and the 17bp DNA helix respectively. The calculated average free energy of each term can be broken down according to the following equation:

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{SOL} - T\Delta S \quad (5.3)$$

$$\Delta G_{bind} = (\Delta E_{int} + \Delta E_{ele} + \Delta E_{vdw}) + (\Delta G_{PB/GB} + \Delta G_{SA}) - T\Delta S \quad (5.4)$$

where the average molecular mechanics energy (ΔE_{MM}) term is a sum of the internal energy (bonds, angles and dihedrals), electrostatic energy and van der Waals term, while the ΔG_{SOL} term accounts for the solvation energy, that comprises both polar and non-polar component. The polar part of the solvation term then accounts for the electrostatic contribution to solvation, and was calculated using a theoretically more rigorous Poisson-Boltzmann (PB) model, and an alternative, computationally-efficient Generalized-Born (GB) model developed by Onufriev *et al*[234] (igb=2; model GB^{OBC1}) with rescaled effective Born radii, accounting for interstitional spaces between atom spheres. For both GB and PB calculations the value of the exterior dielectric constant was set to 80 at 300K, while the solute dielectric constant was set to 1; salt concentration was set to physiological conditions (and corresponding to the MD simulations protocol) ~150 mM. The entropy contributions (TΔS) were neglected in these calculations, since a comparison of two very similar systems, phosphorylated and unphosphorylated STAT3 complexes, was carried out.

The MM/PB(GB)SA calculations were performed employing the last 40 ns (1600 frames, 25-ps/step) of a single trajectory obtained from the explicit solvent MD simulation of pSTAT3βtc-DNA and uSTAT3βtc-DNA complexes, from which the unbound receptor and ligand structures were extracted. Since the two MD simulations were carried out with GROMACS MD package, employing the AMBER force field, the compressed trajectory files (containing the coordinate, time and box vector information; 'xtc' file) needed to be converted into AMBER trajectory file format ('mdcrd' file) for the MM/PB(GB)SA calculations. This was achieved with the VMD visualization program, by uploading the trajectory and saving it in NAMD trajectory file format ('dcd' file), which was subsequently processed by the AMBER coordinate/trajectory processing program *ptraj*, to produce a trajectory in the desired 'mdcrd' file format.

Topologies and input files necessary for the AMBER molecular mechanics programs were generated by means of the LEaP program, via *xleap*, a window-based interface to LEAP. Firstly, a reference pdb file of the pSTAT3 complex and uSTAT3 complex, which represented the conformation at the start of the MD simulation was generated, and subsequently coordinates corresponding to the STAT3:STAT3 homodimer (residues 136 to 716), monomer-A, monomer-B and *ds*DNA helix were saved as separate entities, for both phosphorylated and unphosphorylated systems; those were then processed by *xleap*, providing molecular topology files with the necessary force field parameters stored in them. Corresponding force fields that were used for the MD simulation in GROMACS were employed also here, namely the AMBER parm99sb-ILDN[166] force field, together with the *parmbsc0*[161] force field, and the parameters for the phosphorylated tyrosine residue (in case of the pSTAT3 simulation), were loaded manually.

5.3.6 3D-pharmacophore modeling

The pharmacophore concept, and its simplicity, enable the complexity of intermolecular interactions between ligand and its protein receptor to be reduced to a small set of attributes,[235] such as hydrogen bond donor, acceptor, hydrophobic interaction, or excluded volumes. The LigandScout v 3.0 application framework[236,237] was used to detect crucial interaction patterns within the PPI contact area of pY-containing tetrapeptide PpYLK (residues 704-707) with the SH2 domain of the opposite pSTAT3βtc homodimer-forming monomer.

Here, the 'rigid' 3D structure-based pharmacophore modeling approach was combined with 'dynamic' information with respect to the conformational flexibility of the protein (SH2 domain) and its binding partner (pY-containing tetrapeptide), as obtained from the MD simulations by cluster analysis over the 50-ns time frame. Five representative structures, i.e the middle structures of the five largest clusters sampled at 300 K over the 50-ns MD simulation, were used for the detection of protein contacts via structure-based pharmacophore modeling. The SH2 domain of monomer-A (residues 586 - 688) was always extracted from the representative structures and used as a receptor, while the terapeptide PpYLK (residues 704-707) of the opposite binding partner was extracted, saved as a new pdb entity, and used as a ligand.

Default settings of LigandScout[236,237] were used for ligand interpretation followed by pharmacophore generation. The pharmacophore features and cutoff thresholds were specifically defined within LigandScout as follows:

• The protein-ligand interaction cutoff thresholds were defined in terms of spheres surrounding each non-hydrogen atom of the ligand, and all non-hydrogen atoms of the protein that are within that sphere "environment", whose cutoff distance is 7 Å, were considered to be potential interaction partners.
• In terms of the steric constraints and circumstances of the macromolecule, the minimum and maximum distances were set to 2 Å and 4 Å respectively, such that an excluded volume feature on an alpha carbon was generated if the alpha carbon had in-

teractions with the ligand atoms closer than the maximum distance, and no other feature was closer than the minimum distance.[237]

• Flexible hydrogen bonds were discriminated against rigid hydrogen bonds interactions, relative to the hybridization of their heavy atoms (Figure 5.3); While rigid hydrogen bond interactions, as typically occurring at the sp^2 hybridized heavy atoms, must fulfill the 50° angle range for sp^2 hybridized heavy atoms, the flexible hydrogen bonds, as occurring at sp^3 hybridized heavy atoms, have a 34° default angle range.

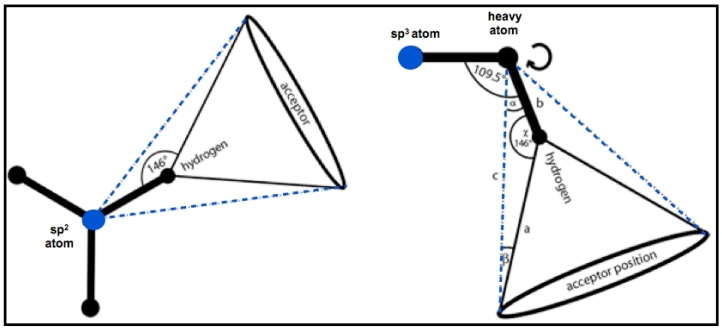


Figure 5.3: Rigid and flexible hydrogen bond constraints on sp^2 and sp^3 hybridized heavy atoms[7].
Rigid hydrogen bond donor represents for instance an sp^2 hybridized amide nitrogen, and flexible hydrogen bond can be found at an sp^3 hybridized hydroxy group.

• Charge interaction features, as well as hydrophobicity features were regarded as distance constraints (the pharmacophore feature definitions and distance constraints parameters outlined in Table 5.2, within the Results section). Similarly, aromatic features which represent π-π (orthogonal and parallel) and cation-π interactions were defined by distance constraints, and also angle constraint between the interacting planes.

[7] Figure adapted, and modified, from Inte:ligand LiganScout v.3 online manual www.inteligand.com/ligandscout3/

5.4 RESULTS AND DISCUSSION

The core of this chapter is focussed on the PPI of STAT3, specifically at phosphotyrosine (pY705) binding in particular, and compared with the unphosphorylated form (Y705). The finding of a significant difference in the shape of pY705/Y705 binding pockets has been incorporated into the molecular docking study of STAT3 SH2 domains with a library of structurally different small molecules (ESP library). Then 3D-pharmacophore modeling of the chemical features characterizing the pY705 binding site over a 50-ns time frame was used to confirm the pY705 binding site stability. Furthermore, thermodynamic stability of the STAT3-STAT3 interaction was evaluated by means of binding energy calculations of both pSTAT3 and uSTAT3 complexes, at the level of PPIs but also reflecting the protein-DNA interactions. Representative structures of the molecular targets were obtained from trajectories of two (25-ns and 50-ns) explicit solvent MD simulations of two forms of STAT3βtc homodimer:DNA complexes, described above.

The results are presented in a comparative manner, in terms of : (1) structural differences of pY705/Y705 binding sites, (2) the molecular docking study of small-molecule ligands with multiple receptor (target) conformations, (3) thermodynamic stability of the STAT3-STAT3 association, and (4) protein-protein contact analysis based on 3D-pharmacophore modeling.

5.4.1 Looking into “the” pocket: structural differences of the pY705 and Y705 binding site

Studies by other research groups have shown that STATs SH2 domains possess strikingly similar backbone conformations that are different from non-STAT SH2 domains[215,238] (which might reflect the high level of structural conservation of the STAT SH2 domains); and that the STAT3 SH2 domain-binding peptide sequence PpYLKTK selectively blocks STAT3 DNA-binding activity, by mechanistically disrupting STAT3:STAT3 dimers (*in vitro*).[53] It has also been shown that the inhibition can be sig-

nificantly improved by choosing phosphopeptide sequences of other STAT3-interacting proteins (such as gp130) and their constrained hybrids.[54,241] Furthermore, it has been shown that modulations of the pY+1 (L706) residue reduced the phosphopeptides binding affinity to the STAT3 SH2 domain, supporting the importance of that particular residue for STAT3 binding.[215] The phosphopeptide binding surface residues (forming three distinct clefts) are known. However, targeting STAT3 domain in order to inhibit protein functions for drug discovery has remained a daunting task.[239]

With the mounting body of evidence that unphosphorylated STAT3 form homodimers (via reciprocal interaction of their SH2 domains) which are able to bind their target DNA sequence, it is appropriate to examine the structural properties of the unphosphorylated STA3 SH2 domain binding site with respect to the phosphorylated one. This information can be then used in molecular docking studies with therapeutic small-molecules. Dynamic point of view into the phospho- and unphosphorylated-binding pockets, defined by a pentapeptide sequence P-pY/Y-L-K-T, has been employed here. Three representative structures from the respective 25-ns MD simulations were used, corresponding to ~70% of the conformational space sampled at 300K. All six representative structures of the SH2 domains with the pentapeptide of the opposite STAT3 monomer (three phospho- and three unphospho-STAT3 complex bound monomers) were visually explored. A distinct feature of the pY705/Y705 -accommodating pocket has been observed (Figure 5.4); in the case of the pY705, residues K591, R609, S611, E612 and S613 form a tight clamp-like binding pocket around the polar (negatively charged) phosphate group (Figure 5.4 a), however in the case of the unphosphorylated Y705 residue, the shape of the binding pocket is significantly changed due to a rearrangement of residues K591 and E594 (Figure 5.4 b). Residue K591, which would normally interact with pY705, and contribute to the specific shape of the pY-binding site, is now completely flipped out of the site, with its position being partially replaced by E594. It can be speculated that this is a consequence of the relatively bulky and negative charge-carrying phosphate group being replaced by a hydroxyl group from the regular tyrosine residue, which does not attract the lysine residue (with protonated side chain at simulated neutral condition) so strongly.

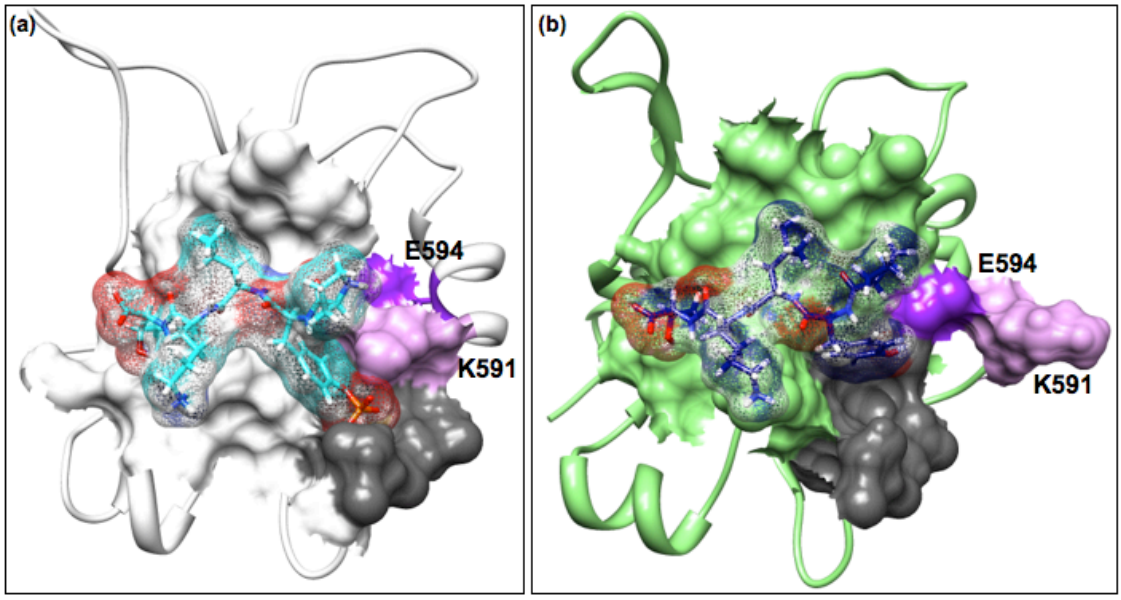


Figure 5.4: SH2 domain binding pocket of the Y705/pY705-containing pentapeptide (res 704-708). (a) phosphorylated PpYLKT pentapeptide bound to the opposite dimer-forming SH2 domain; (b) unphosphorylated PYLKT pentapeptide bound at the SH2 binding site of the opposite unphosphorylated STAT3 monomer. Residues within 7Å of the pentapeptide binding site are used in molecular surface representation (*white/light green*), while the pentapeptide is shown in licorice representation (*cyan/dark blue*). Residues R609, S611, E612 and S613 are shown in *gray* while the fifth, pY705 binding pocket-forming residue K591 is displayed in *lilla*, and E594 in *purple.*

The altered, wider, shape of the Y705 pocket with respect to the more enclosed pocket formed around pY705 was observed for all three unphosphorylated target representations that were explored and used for subsequent comparative molecular docking studies, described in the next section. With regard to the very different shape of the Y705/pY705 binding pocket, resulting from a phosphate group presence/absence, it might be expected that structurally different ligands will preferentially bind there. The chemical environment of the pocket, where either the positively charged lysine, or un-protonated glutamate comes into direct interaction with a pocket-occupying ligand may affect the preference for a suitable ligand. This knowledge has to be kept in mind with respect to the subsequent molecular docking study of potential STAT3 inhibitors, since both phosphorylated and unphosphorylated STAT3 dimers, and subsequently STAT3-DNA complexes can occur.

5.4.2 Molecular docking with multiple target conformation of the STAT3 SH2 domain

54 chemically and structurally diverse small molecules, that were selected by medium throughput screening of ~25.000 compounds (ESP Laboratories Hamburg) against the unphosphorylated STAT3 SH2 domain with phosphorylated hexapeptide probe, were used for a multiple-target conformation molecular docking study, using the two well-established programs for protein docking, GOLD and DOCK6. The molecular target representations were obtained by cluster analysis of 25-ns trajectories of the pSTAT3-DNA and uSTAT3-DNA complexes respectively. In both cases, the three middle structures of the three largest clusters spanned over ~70% of the conformational space (sampled at 300K), and were used as molecular targets in the docking study. The consistency and comparability of the results was then secured by using a total of six corresponding conformations of the STAT3β SH2 domain in both GOLD and DOCK6 small-molecules docking studies. The general overview of the comparative molecular docking study that was applied here is shown in Figure 5.5.

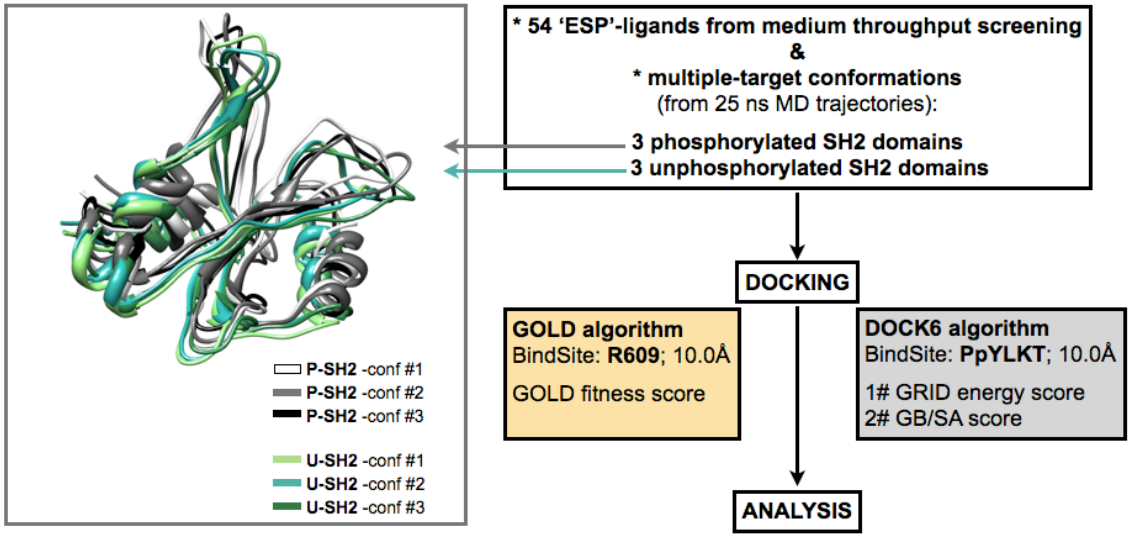


Figure 5.5: The overall outline of the molecular docking study.
A total of six rigid conformations of the molecular target of STAT3β-SH2 domain (phosphorylated/ unphosphorylated) and two molecular docking programs, GOLD and DOCK were employed; The structural alignment is shown for the target representations - SH2 domains from phosphorylated (*grey shades*) and un-phosphorylated (*green shades)* STAT3:DNA complex, as obtained by cluster analysis of the 25-ns MD trajectories.

Whereas the default optimized setting was used while employing GOLD, DOCK6 required careful selection of the input parameters in order to obtain sensible binding poses of the small molecules within the target binding site. Parameter selection was tested and optimized based on the "trial-and-error" method, with the final settings validated by docking pY705 and Y705 residues in their respective binding sites within the three STAT3 SH2 domain conformations (Figure 5.6; Table S5.1 in the Supplementary section). Docked pY705 with pSTAT3 SH2 domain conformations 1 and 2 displayed excellent agreement in terms of the binding pose and orientation with the phosphopeptide-bound pY705 (Figure 5.6 a and b). In the third case, the phosphate and aromatic parts of pY705 were overlapping between the docked and phosphopeptide-bound residues, but the amido-portion of the residue was observed to have partially flipped (Figure 5.6 c). A well-defined pY705-binding pocket, as well as the characteristic L706 binding pocket, were observed for all three conformations of the pSTAT3 SH2 domain.

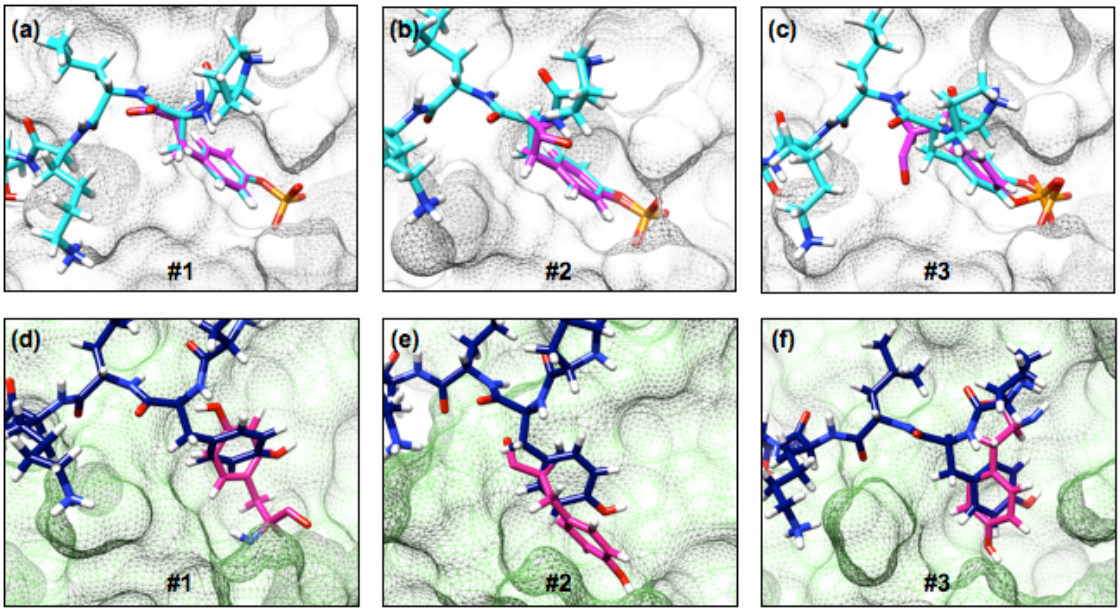


Figure 5.6: pY705 and Y705 docked into their respective binding sites
pY705 and Y705 were docked with three receptor conformations (i.e six in total) as obtained from the MD runs of pSTAT3 and uSTAT3 complexes; (a-c) pY705 docked with three conformations of the opposite STAT3 SH2 domain; (d-f) Y705 docked with three conformations of the SH2 domain, its binding partner from the uSTAT3 complex. SH2 domains are shown in surface mesh representation (*white/ pale green*), while the binding site-defining pentapeptides of the opposite SH2 domain are shown in stick representation (*cyan/ navy blue*). Docked pY705 and Y705 are colored in *magenta* and *pink* respectively.

On the other hand, Y705 docked with the pentapeptide binding site of the uSTAT3 SH2 domains did not display a strong correlation with the pentapeptide-bound Y705 poses and orientations, when the corresponding docking parameters were used (Figure 5.5 d-f). The first and the third target binding site conformations had slightly more ordered structural features than the second conformation. This may explain the greater overlap of the tyrosine aromatic-ring portions that was observed for the otherwise weakly overlapping Y705 residues. However, the shallower and more open binding site of uSTAT3 SH2 domain provides a less-restricted conformational space for the ligand to explore and adapt to, as is visible even from the different orientations of pentapeptide-bound Y705.

During the docking of the 54 small molecules with total of six different representations of STAT3 SH2 domain (using GOLD and DOCK6), the ligands predicted binding poses within previously defined binding sites were scored by means of (1) the dimensionless fitness function, GoldScore[223](GOLD), which is optimized for the prediction of ligand binding positions rather than predictions of binding affinities; and (2) the Hawkins GB/SA[230,231] scoring function which was employed in the final minimization and ligand scoring/ranking routine in DOCK6, which provides ligands binding affinities for the molecular target in kcal/mol. The results obtained were filtered for subsequent analysis according to the following criteria:

• First, ligands that did not successfully dock with all six target conformations (i.e three phospho- and three unphosphorylated STAT3 SH2 domain representations) in both GOLD and DOCK6-employing docking studies were removed (a total of 12 compounds).

• Second, mean values (and their standard deviations) of ligand fitness scores (GoldScore) and binding affinities (secondary GB/SA score, primary GRID-based score) were calculated over the three conformations of the phosphorylated, and unphosphorylated molecular target, for both GOLD and DOCK6 predictions. Ligands whose calculated mean fitness score/binding affinity for the phosphorylated molecular target (p-SH2 domain) had high values (stdev ≥ 10) their respective standard deviations were removed (a total of three compounds).

• Third, ligands whose mean fitness score/binding affinity values were smaller than half of the top ranked compound by means of GoldScore and GB/SA scores within both phosphorylated and unphosphorylated STAT3 SH2 domains (p-SH2 and u-SH2) were removed. For example, if the top mean value of a compounds fitness score with the p-SH2 target was estimated as 44 by means of GoldScore, those ligands with mean fitness score smaller than 22 were removed (a total of six compounds).

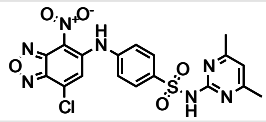
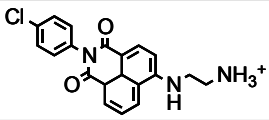
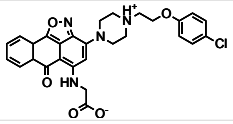
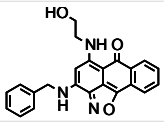
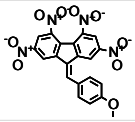
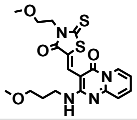
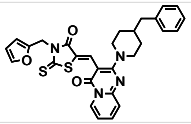
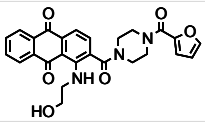
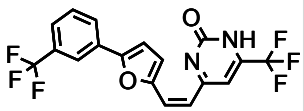
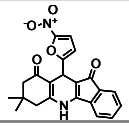
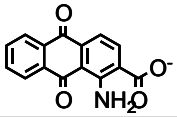
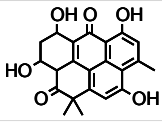
34 compounds remained for further analysis upon processing the docking results through the "criteria selection funnel". Based on fitness scores/binding affinities of the ligands, the top six ligands were selected for the phosphorylated and unphosphorylated molecular targets (p-SH2 and u-SH2) (Table 5.1). These compounds, presented in four categories (i.e GoldScore p-SH2, GoldScore u-SH2, GB/SA secondary score p-SH2 and GB/SA secondary score u-SH2) were combined, leading to a total of 10 ligands from the original ESP-library of chemically-divers (CD) small molecules.

Table 5.1: Top six compounds predicted by MRC docking with GOLD and DOCK6.
GoldScore (GOLD) and GB/SA binding affinity secondary scoring function (DOCK6) for phosphorylated and unphosphorylated receptor (STAT3 SH2domain). Ligands with experimentally-determined optimal dose response curve are marked in blue.

Rank	GOLD		DOCK6	
	GoldScore; p-SH2	GoldScore; u-SH2	GB/SA score; p-SH2	GB/SA score; u-SH2
1	26810_CD	25113_CD	29768_CD	28373_CD
2	25113_CD	24601_CD	28373_CD	24601_CD
3	28373_CD	26810_CD	26810_CD	26745_CD
4	25117_CD	25117_CD	24601_CD	26810_CD
5	29768_CD	24846_CD	26745_CD	25113_CD
6	29292_CD	28373_CD	25117_CD	28018_CD

Among the combined top 10 (Table 5.1 and 5.2) potential small-molecule inhibitors, ligands 26810_CD and 24601_CD were shown to have optimal dose response curve shapes (determined experimentally, prior to the start of the computational study, Figure 5.1). Furthermore, ligand 26810_CD and 28373_CD were selected among the top six compounds in each of the four categories; ligands 24601_CD, 25113_CD and 25117_CD occurred among the top six compounds three times; ligands 29768_CD and 26745_CD twice; and ligands 29292_CD, 280180_CD and 24846_CD once. When comparing only predictions for the phosphorylated target, there is a match in four out of the top six compound between GOLD and DOCK6, which is 75% agreement.

Table 5.2: The 10+2 top compounds selected via MRC docking (p-SH2 and u-SH2).
The top 10 compounds in the table are the combined best scored compounds as listed in Table 5.1; the two compounds at the bottom (highlighted in *grey*) were additionally analyzed due to their optimal shape of dose-response curves that were previously experimentally determined. The compounds are listed in a numeric order. GoldScore

Compound id #	Structure	GOLD		DOCK6	
		p-SH2	u-SH2	p-SH2	u-SH2
		GoldScore	GoldScore	GB/SA score [kcal/mol]	GB/SA score [kcal/mol]
26810_CD		58.89 ± 3.73	53.34 ± 5.27	-27.37 ± 2.83	-23.68 ± 7.46
24601_CD		39.46 ± 7.44	54.44 ± 10.62	-27.14 ± 4.52	-25.74 ± 2.53
25113_CD		55.90 ± 5.53	55.34 ± 0.35	-26.21 ± 2.11	-22.73 ± 2.71
25117_CD		49.77 ± 2.94	53.13 ± 2.27	-26.39 ± 1.76	-18.82 ± 10.36
26745_CD		46.24 ± 6.46	47.32 ± 3.82	-27.12 ± 0.52	-24.11 ± 6.15
28373_CD		51.54 ± 1.43	52.05 ± 6.02	-30.17 ± 2.75	-27.05 ± 4.36
29768_CD		49.26 ± 6.58	48.26 ± 3.44	-30.47 ± 1.03	-21.19 ± 14.24
24846_CD		43.03 ± 3.09	52.38 ± 4.75	-24.28 ± 1.94	-18.05 ± 12.53
28018_CD		38.05 ± 4.75	41.91 ± 2.98	-23.52 ± 1.85	-22.62 ± 2.22
29292_CD		47.06 ± 3.22	46.26 ± 8.20	-23.44 ± 0.25	-20.02 ± 3.50
24712_CD		46.49 ± 8.34	47.12 ± 3.45	-19.21 ± 4.14	-13.36 ± 4.79
28580_CD		40.20 ± 2.74	45.60 ± 5.12	-19.41 ± 4.47	-14.75 ± 0.55

Considering the experimental data for all 54 ligands, and in particular those with optimal shape of their dose-response curves, ligands 24712_CD and 28580_CD were further analyzed together with the top 10 selected ligands, as they passed the initial "criteria selection funnel", but failed to be among the top six compounds in four categories. Hydrogen bonds formed between the each of the 12 (10+2) ligands docked with a total of six target conformations were calculated by means of UCSF Chimera program (Table 5.3) and compared.

Table 5.3: Hydrogen bond-forming residues of the MRC docked with selected ligands.
Hydrogen bonds formed between the top 10+2 ligands previously selected (Table 5.2) and p-SH2/u-SH2 receptor representations (six conformations in total; 3+3) were determined by means of UCSF Chimera program. In the table, three lines per ligand correspond to three conformations examined. Protein residues involved in hydrogen bond interactions are listed for each of the three receptor conformations.

Ligand id #	p-SH2 hydrogen bond-forming residues		u-SH2 hydrogen bond-forming residues	
	GOLD p-SH2	DOCK p-SH2	GOLD u-SH2	DOCK u-SH2
24601	K591	E625	I589, E594, E612, S636	E594
	Q635, S636, T620	K591, E638	E594, R595, S636	E625
	E594, I634, S636	E625	E594, I634, S636	S636, E638
24712	R609, S611, E612, S613	R609, S611, E612, S613	S611, E612, S636	S611, E612, S636
	R609, S611, E612, S613	Q633	R595	R595
	E592, R609, E612	R609, E612	I589, R609, F610	K591
24846	E612	---	E594, R609	S636
	K591, R595, E638	K591, E594, R609, S636	R595	S611, S636
	S611. E612	---	R609, T620	S636
25113	R609, E612	Q635	S611, E612, S636	---
	R609	K591, S636	R609, E612, S613, S636	S611
	R609, E612, S613	R609, S611, E612, S613	K591	---
25117	Q635	S611	E612, V637	---
	R609, S636	E594, S636	E594, R609, E612, S613	E594, S611, S613, S636
	R595, R609	M660	E594, R609, S611	---
26745	K591, R609, Q635	Q644	S611, E612, S636	S611
	K591, E592, R609, E612	Y657	S611, S613	S611
	K591, R609, E612	---	R609	S611
26810	R609, S611, E612, S613	S636, E638	E594, S611, E612, S613	R609, E612, S613
	R595, R609, S611, E612, S613	S636	R595, I634	S611, E612
	R591, R595, R609, S611, E612, S613	S636	I589, I634	K591
28018	---	K591, R609, S613	S636	S611, E612
	R609	K591, S613	K591, R595	S636
	K591, R609, E612	R609, E612	R595	R595
28373	Q635	Q635, S636, E638, Y657	K591, S611, S636	S636
	R595	K591	K591, E612, S613	E594, R595
	K591, R609, E612	K591, E612, S613	E594, R609, S611, E612, I634	---
28580	S611, S613, E638	---	T620, S636, V637	E594, V637, E638
	K591, S636, V637	R595, Q633	E612, S613, I634, S636	E594, S613, S636
	R609, S611	I659, M660	E592, E594, R595, I634	---
29292	K591, R609, S611, E612, S613	K591, R609, S611, S613	E594, S611, E612	S611
	R609, S611, E612, S613, E638	R595, S636	R595	---
	R609, S611, E612, S613	I659, K658	S613	K591, S611, S613
29768	K591, R609, E612	K591, R609	R593	---
	K591, R609, S611, E612	K591, R609	---	R595
	E592, R609, E612	K591, R609	R595, I634	---

The analysis of hydrogen bonds formed between the ligands docked with MRC of the pSTAT3 and/or uSTAT3 SH2 domain can be summarized as follows:

• more hydrogen bonds were predicted to be formed between ligands docked with the p-SH2 domain than with u-SH2 domain conformations by ~ 25%, which may reflect the tighter, more enclosed shapes of the p-SH2 receptor pockets, comparing to the wider and shallower pockets formed by u-SH2.

• in terms of the ligands docked with the p-SH2 MRC, the total number of hydrogen bonds determined within GOLD-predicted interactions (101 hydrogen bonds) is ~ 55% higher than for DOCK6-predicted solutions (65 hydrogen bonds), suggesting that GOLD-predicted binding poses my be more accurate predictions.

• the difference in the total number of predicted hydrogen bonds by GOLD and DOCK6 is even higher for the ligands docked with the u-SH2 MRC; as the GOLD-predicted binding poses (87 hydrogen bonds) of the ligands formed ~ 93% more hydrogen bonds than the DOCK6-predicted (45 hydrogen bonds) binding poses.

• those protein residues forming hydrogen bonds with the ligands upon docking, were generally in better agreement across the three receptor conformations of p-SH2 domain than for the u-SH2 domain. A better correspondence was also found among the GOLD-suggested solutions when compared to the DOCK6 solutions (i.e ligand 24712_CD, 26745_CD, 26810_CD, 29292_CD, and 29768_CD);

The ligands binding poses and orientations within the MRC pocket were further examined, with the following findings:

• in a number of cases, the GOLD-predicted binding poses of a flexible ligand were targeting a corresponding pocket, with at least two of the three ligand orientations in agreement, but the DOCK6-predicted binding poses were spread all over the initially defined binding site (i.e 26810_CD; Figure 5.7 a-d).

• for ligand 29768_CD, both GOLD and DOCK6 ligand binding poses and orientations were in good correspondence among the p-SH2 domain receptor conformations, but quite diverse for the u-SH2 (Figure 5.7 e-h).

• also the opposite situation, where ligand binding poses predictions were barely correlated for the p-SH2 receptor conformations in both GOLD and DOCK6, but well-matched for the u-SH2 receptor was found for ligand 25117_CD (Figure 5.7 i-l).

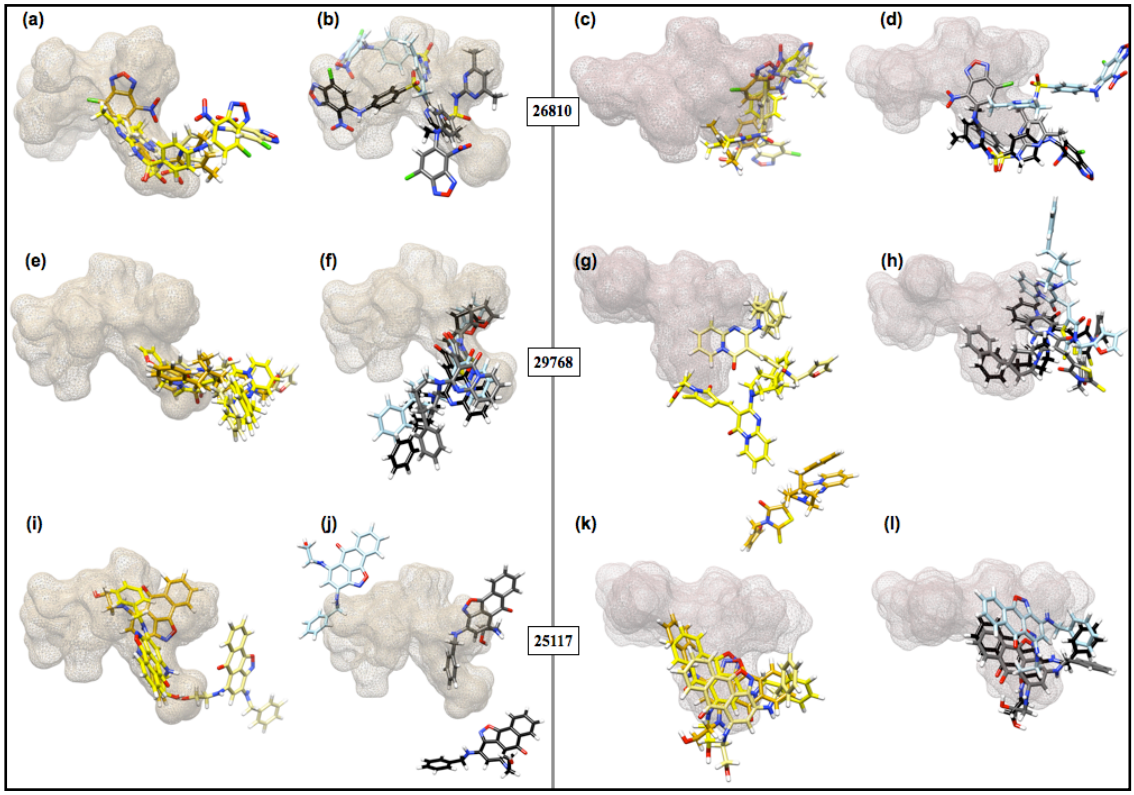


Figure 5.7: Structurally-overlaid selected ligands docked with MRC of p-SH2 and u-SH2 domain. Each line in the panel shows predicted binding poses of the ligands with p-SH2 and u-SH2 receptor conformations (a total of six) as obtained by means of GOLD (*gold, yellow* and *pale yellow*) and DOCK6 (*black, gray* and *pale blue*). The defined binding sites are shown in molecular surface mesh representation, *beige* for three p-SH2 domain conformations (on the left side of the panel) and *pale pink* for three u-SH2 domain representations (on the right side of the panel). All ligands are shown in licorice representation; (a-d) ligand 26810_CD, (e-h) ligand 29768_CD and (i-l) ligand 25117_CD.

The first target representations (i.e the most populated conformation found in the 25-ns MD simulation) for both p-STAT3 and u-STAT3 SH2 domains were further considered for the comparison of ligand poses and orientations predicted by GOLD and DOCK6. Excellent agreement was observed for two relatively rigid ligands 24712_CD and 29292_CD docked with the p-SH2, and they also displayed a good correlation within the u-SH2 binding pocket (Figure 5.8 a-b, c-d). In contrast, a flexible ligand 26745_CD, that has seven rotatable bonds, showed excellent correlation between DOCK6 and GOLD binding pose prediction in the u-SH2 target, but not in p-SH2 (Figure 5.8 e-f). This suggests that the choice of a docking software may have a significant impact on the selection of a promising inhibitor candidate. Combining the DOCK6 and

GOLD predictions together with the MRC that were examined, a potent STAT3 inhibitor would be expected to preferentially occupy the same binding pocket of both phosphorylated and unphosphorylated receptor, with a decent agreement of both software tools employed. Accordingly, rigid ligands such as 24712_CD and 29292_CD, or quite contrary, flexible ligands such as 29768_CD or 26810_CD were found to fulfill these criteria.

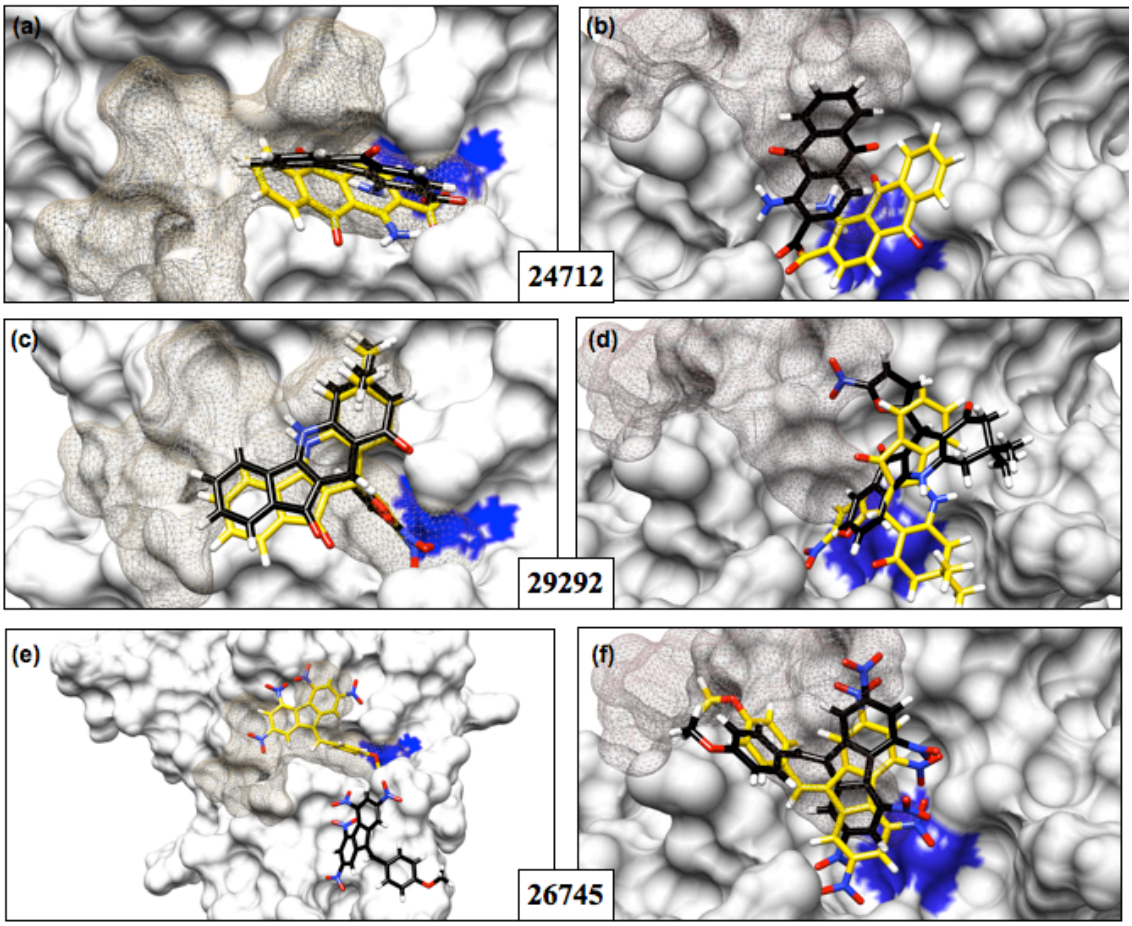


Figure 5.8: Comparison of the GOLD and DOCK6 predicted ligand's binding poses.
Selected ligands are docked with the most populated conformation of p-SH2/u-SH2 domain. The p-SH2 domain (*white*) and the u-SH2 domain (*grey*) are shown in solid molecular surface representation. The space occupied by the PpY/YLKT pentapeptide (res 704-708) is shown in mesh molecular surface representation in *beige* and *pale pink* color respectively. The position of R609 within the molecular target is highlighted in blue (as a reference point). Binding poses of the individual ligands are always shown in licorice representations; carbon atoms are shown in *gold* (GOLD) and *black* (DOCK6), and further color-coded by atom name (oxygens in *red*, nitrogens in *blue*, hydrogens in *white*). (a-b) ligand 24712_CD, (c-d) ligand 29292_CD, and (e-f) ligand 26745_CD.

5.4.3 Binding free energies of the STAT3:STAT3 and STAT3 dimer:DNA intermolecular association calculated by means of the MM/PB(GB)SA method

The interaction energy and the solvation free energy for the phosphorylated (1) and unphosphorylated (2) STAT3:STAT3 (protein-protein) intermolecular association, as well as for the phosphorylated (3) and unphosphorylated (4) STAT3 homodimer:DNA complex formation was calculated from the explicit solvent MD trajectories of the respective STAT3 complexes (Figure 5.9), by means of the post-processing MM/PB(GB)SA method as implemented in MMPBSA.py in AMBER 11. The single trajectory approach was employed, hence configurational ensembles for the complex (STAT3-DNA complex or STAT3:STAT3 homodimer), receptor (STAT3:STAT3 homodimer or STAT3 monomer-A) and the ligand (*ds*DNA helix or STAT3 monomer-B) were extracted from a single MD trajectory of the phosphorylated and unphosphorylated STAT3βtc-DNA complex respectively. The choice of this method was convenient because of its faster calculation time, comparing to the multiple trajectory approach. Also the generation of separate trajectories via explicit solvent MD simulations of all complex-forming components would not be feasible due to the computational time required for this size of the simulated systems (~1235 residues). However, one needs to be aware of this method's assumption that there is no significant structural or dynamic changes between the bound and unbound macromolecules, potentially leading to incorrectly estimated ΔG values.[135,241]

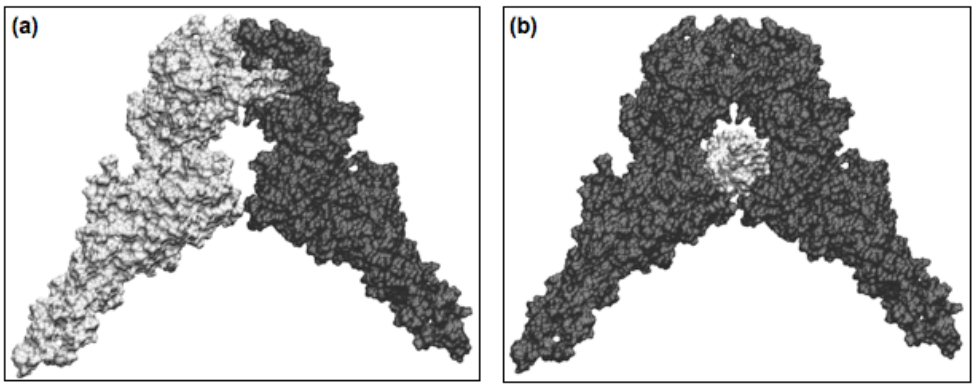


Figure 5.9: Representation of the intermolecular association of the STAT3 complex formation.
(a) shows the STAT3:STAT3 association, where the STAT3 monomer-A (*black*) was regarded as a receptor for the calculation, while monomer-B (*silver*) was assigned as a ligand; (b) represents the STAT3-DNA complex (protein-DNA) intermolecular association, with STAT3 homodimer (*black*) being the receptor, and the *ds*DNA (*silver*) is the ligand. The corresponding principle was applied both to the pSTAT3 complex and uSTAT3 complex.

The importance of entropic contributions to the otherwise "enthalpic view" of free binding energy of macromolecular association has been emphasized by numerous studies, and reviewed for instance by Harris and Laughton.[133,151] The conformational entropy calculations have proven to be very challenging, and the inclusion of an entropy term has been shown to not always improve prediction accuracy.[136] Moreover, since the free energy calculations of very closely related systems were carried out here, the only difference between the pSTAT3 and uSTAT3 complexes being the phosphorylation of incremented residue Y705, it can be assumed that the entropic contributions to the absolute binding free energies would cancel each other out when we are interested in the relative free energy of binding between a phosphorylated and unphosphorylated complex, either at the protein-protein or protein-DNA level of biomolecular association.[135,229]

The calculated binding free energy (ΔG_{bind}) for the STAT3-STAT3 interaction (PPI) in the pSTAT3βtc complex was -201.36 kcal/mol using the MM/PBSA method, and -186.72 kcal/mol using the MM/GBSA method, while -102.14 kcal/mol (MM/PBSA) and -99.56 kcal/mol (MM/GBSA) respectively, for the uSTAT3 protein-protein interaction. In terms of protein-DNA association (PDI), the corresponding calculated binding energies for the pSTAT3βtc and uSTAT3βtc complexes are: ΔG_{bind} values of -140.10 kcal/mol (MM/PBSA) and -96.17 (MM/GBSA) for the pSTAT3:DNA complex, and -140.73 kcal/mol (MM/PBSA) and -97.53 kcal/mol (MM/GBSA) for the uSTAT3:DNA complex. Both PB and GB methods predicted that the binding free energy of the protein-protein association is two-fold more favourable for pSTAT3βtc than for the uSTAT3βtc complex. These energies (summarized in Table 5.4) are in good qualitative agreement with the experimental data described by Nkansah *et al.* (manuscript submitted for publication, 2012). By employing combined experimental techniques (PEMSA and x-ray crystallography), uSTAT3 protein binding to *ds*DNA was demonstrated, and its strong similarity to pSTAT3-DNA interaction was shown. These observations have been further supported by CD spectroscopy, that suggested well-folded and stable pSTAT3-DNA and uSTAT3-DNA complex conformations, and by MD simulations described above.

Table 5.4: Overview of the MM/PB(GB)SA binding energies for pSTAT3 and uSTAT3 protein-protein and protein-DNA interactions.

	ΔG_{bind} [kcal/mol]		pSTAT3 : uSTAT3
	MM/PBSA	**MM/GBSA**	**ratio**
pSTAT3 - pSTAT3	-201.36	-186.72	2:1
uSTAT3 - uSTAT3	-102.14	-99.56	
uSTAT3 - DNA	-140.10	-96.17	1:1
pSTAT3 - DNA	-140.73	-97.53	

While there is a good agreement between the PB and GB methods at the PPI level for the uSTAT3 complex (~3 kcal/mol difference), the presence/absence of pY705 in the calculated system caused a difference of additional ~10 kcal/mol between the PB and GB method binding energy predictions. The difference between predicted biding energies by means of PB and GB methods is, however, significantly larger for the PDIs, as the calculated difference between the two methods is ~45 kcal/mol for both pSTAT3-DNA and uSTAT3-DNA complexes. This is not too surprising though, as the accuracy of PDIs binding energies by means of MM/GBSA methods is known to be lower than for PPI complexes. Furthermore, the ratio of the predicted binding energies (ΔG_{bind}) remained constant (2:1 and 1:1 respectively) for both MM/PBSA and MM/GBSA calculations, providing the aimed-for estimate of free energy difference of the STAT3 binding with respect to the phosphorylated vs unphosphorylated, protein-protein and protein-DNA association.

5.4.4 Protein-protein contact analysis based on 3D-pharmacophore modeling

Experimental studies of highly conserved SH2 domains have revealed that ~50 % of the binding affinity is attributed to the phosphate moiety of the pY705 residue itself, while residues in positions from -2 to +4, relative to pY705, modulate binding specificity.[214] In particular pY705 and L706 have been shown (by means of SAR, and alanine scanning mutagenesis studies) to be critical for STAT3 binding. Therefore multiple representative structures of the pY705-containing tetrapeptide PpYLK (residues 704-707) binding to the SH2 domain of the opposite pSTAT3 homodimer-forming monomer-A were employed in this 3D structure-based pharmacophore modeling

approach. This allowed mapping out (and confirmation) of the characteristic features of the interaction to be undertaken, while utilizing the information of the protein conformational flexibility as suggested by MD simulation. With respect to the principal focus being at the protein-protein interface, only the SH2 domains with the stretch of C-terminal residues were considered for the RMSd-based cluster analysis in this case. Five representative structures obtained from the cluster analysis of the 50-ns trajectory (48 ns respectively) of the pSTAT3-DNA complex, spanned over ~ 86% conformational space sampled, with the first two clusters representing 58% (32% + 26%).

The hypothesis obtained 30 features, summarized in Table 5.5 and graphically shown in Figure 5.10.

Table 5.5: Specific features of the PpYLK-SH2 domain interaction represented by 3D-pharmacophores. Specific features observed at the five representative structures are all summarized here.

Pharmacophore feature		distance constraints	tetrapeptide PpYLK	receptor residues
Hydrogen bond donor		2.2 - 3.8 Å	L706(NH)	S636
			K707(Nz; ε-amino side chain)	E638
Hydrogen bond acceptor		2.2 - 3.8 Å	pY705(OP)	K591, R609, S611, E612, S613
			pY705(OS)	R609, K591
			L706(O)	E638
			P704(O)	Q635
Hydrophobic interactions		1.0 - 5.9 Å	pY705(aromatic side chain)	T620, P639
			L706(aliphatic side chain)	W623, V637
Negative ionizable area		1.5 - 5.5 Å	pY705(OP)	K591, R609
Positive ionizable area		1.5 - 5.5 Å	K707(Nz; ε-amino side chain)	E638
Excluded volume		sterical circumstances	tetrapeptide PpYLK	K591, R609, S611, E612, S613
				T620, F621, W623, Q635, S636
				V637, E638, P639, Y657

With respect to the PpYLK ‘ligand’, there are two hydrogen bond donors and six acceptors, two hydrophobic groups/interactions with four different residues, one negative and one positive ionizable area, and 14 excluded volumes, which are derived from the sterical circumstances of the macromolecule. The two hydrogen bond donor features of the PpYLK tetrapeptide (residues 704-707) reflect the (i) NH group of L706, which hydrogen bonds with the side chain (OH group) of S636, and (ii) the protonated amino group of the K707 side chain, forming hydrogen bond with the un-protonated side chain of E638. The four hydrogen bond acceptor features of the PpYLK ‘ligand’ then reflect hydrogen bond formation between (i) the negatively charged phosphate group of pY705

with five residues forming a tight pocket where the phosphate group of pY705 binds, namely R609 and K591, S611, E612 and S613 (which is in agreement with other theoretical and experimental studies[216]); (ii) the backbone atoms involved in hydrogen bond formation between L706 (tetrapeptide) and E638, which makes L706 and E638 both hydrogen bond donors and acceptors respectively; and (iii) a hydrogen bond between the carbonyl oxygen of P704 with Q635.

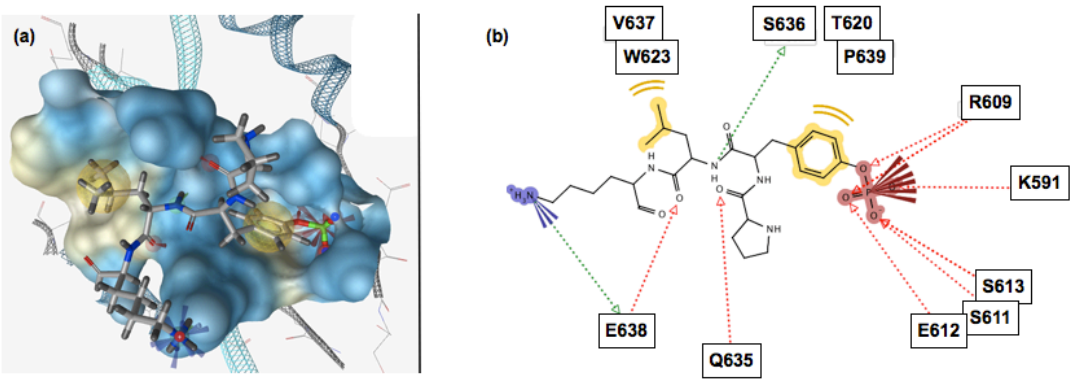


Figure 5.10: Structure-based 3D-pharmacophore model of the PpYLK tetrapeptide.
The model is based on second representative structure obtained from the cluster analysis of the 48 ns MD trajectory. (a) 3D-pharmacophore is shown in licorice representation, with its binding pocket shown as molecular surface. (b) 2D representation of the PpYLK pharmacophore shows the specific features describing the interaction with the SH2 domain.

The charge interaction features, i.e negative and positive ionizable areas, are defined by distance constrains only. The interactions evolve around the charged portions of the tetrapeptide (PpYLK) ligand; specifically at the negatively-charged phosphate group of pY705 interacting with the positively charged side chains of K591 and R609. Also the positive charge-carrying side chain of K707 which forms further electrostatic interaction with the hydrogen-bond forming un-protonated side chain of E638. Hydrophobic spheres were occupied, for all five conformations that were examined, by the aromatic ring of pY705. This pharmacophore feature is reflected in the hydrophobic pocket formed by T620 and P639. Distinct hydrophobic interactions are also formed between the branched side chain of L706 with the receptor residues W623 and V637, which form a relatively shallow hydrophobic pocket. The 14 excluded volume features

of the receptor, as determined by LigandScout over the five representative structures, reflected potential steric restrictions, and correspond to the positions sterically claimed by the macromolecular environment surrounding the pY705-containing tetrapeptide (residues K591, R609, S611, E612, S613, T620, F621, W623, Q635, S636, V637, E638, P639 and Y657). A combined structure-based 3D-pharmacophore model (representing ~60% of the sampled conformational space at 300K) with the hypothesized key chemical featured for the PpYLK tetrapeptide recognition is shown in Figure 5.11.

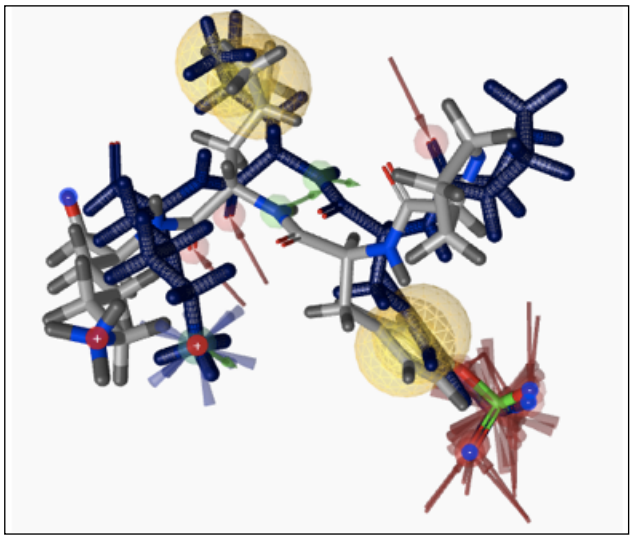


Figure 5.11: Shared 3D-pharmacophore for the PpYLK tetrapeptide.
The shared structure-based pharmacophore model representing the features of the PpYLK-SH2 domain interaction as found for the first two - first (*grey*) and second (*black*) - representative structures. Hydrogen bond interactions are shown as arrows, donors in *green*, acceptors in *red*. Hydrophobic interaction features *yellow* sphere, while the charge interactions are shown as *blue* (positive) and *red* (negative) cones.

The pharmacophore-based analysis of the residues forming the protein-protein contacts in positions from -1 to +3 with respect to pY705 by LigandScout, confirms the importance of the known pY705 pocket forming residues (K591, R609, S611, E612 and S613) and other tetrapeptide-accommodating residues. For instance McMurray[215,241] *et al* reported a detailed description of STAT3 SH2 domain residues involved in direct interaction with high affinity phosphopeptides. Their data were derived from static molecular docking simulations, and two 3-ns MD simulations of the SH2 domain-ligand (constrained phosphopeptides) complexes respectively. Here, the STAT3-own tetrapep-

tide sequence PpYLK was studied utilizing the knowledge of conformational flexibility obtained from the 50-ns MD simulation, thus providing reassuring information about the importance of certain previously reported contact residues, but also demonstrating the features of the network over time. L706 has indeed been shown to be key residue in terms of modulating STAT3:STAT3 specificity, since it acts as both a hydrogen bond donor (with S636) and a hydrogen bond acceptor (with E638) through interactions employing its backbone amid-nitrogen and carbonyl-oxygen respectively. Its branched aliphatic side chain is accommodated in a hydrophobic pocket defined by W623 and V637. Also E638 was shown to be important hydrogen bond donor and acceptor simultaneously, whereas Q635 seems to be in direct, hydrogen-bonding, interaction with P704, at position pY -1, which is only at one of the five studied conformations obtained from MD. This suggests that P704 is perhaps not so important in terms of structural recognition and the binding pocket definition for a potential high-affinity ligand. On the other hand, persistent hydrophobic interaction of the pY705 aromatic ring with T620 and P639, help to define the tight clamp-like pocket around the phosphate portion of pY705, which in turn is wide open toward the following residue L706, thus providing much more space for structural variety of the ligands functional groups (Figure 5.10 a).

5.5 CONCLUSIONS

Conformations of pSTAT3-DNA complex and uSTAT3-DNA complex sampled through two parts of explicit solvent MD simulations were utilized in a multiple-approach *in silico* study of the STAT3 PPI interaction, with the primary focus at the pY705/Y705 binding region. To date, no comparable computational study of this scope or approach, where phosphorylated and unphosphorylated STAT3 forms were compared, and their structural differences further applied in MRC docking study has been reported. It is proposed, that this comparison approach is relevant to potential small-molecule STAT3-STAT3 inhibitor intervention, since a mounting body of evidence has suggested a non-canonical STAT3 signaling pathway, leading to the unphosphorylated STAT3 dimer forming a complex with its target DNA sequence in the nucleus.

(I) Distinct features of the unphosphorylated Y705-binding pocket were found with respect to the widely studied and structurally described pY705-binding site. In multiple target representations K591 was found to be flipped out of the tight pY binding pocket (formed of residues K591, R609, S611, E612 and S613) in the case of the unphosphorylated tyrosine residue, and replaced by E594. This partial pocket re-arrangement leads to a wider, less structurally-defined binding site for the uSTAT3 SH2 domain. The chemical features of the binding pocket are also influenced and/or altered when the K591 was observed to have flipped.

(II) A MRC docking study of 54 experimentally-determined small molecules docked with both phosphorylated and unphosphorylated STAT3 SH2 domain (in a total of six conformations), using two docking softwares, provided a more thorough insight into the ligand-receptor (STAT3 SH2 domain) interactions at numerous levels. SH2 domains, including the pY-binding pocket, are principally known to be structurally conserved. However subtle conformational changes have shown to have a strong effect on predicted binding poses and orientations of small molecule ligands that were examined, and has demonstrated that this combined docking approached provides an advantage over the conventional single receptor conformation concept. Ligands 24712_CD, 29292_CD, 29768_CD or 26810_CD have consistently shown good results in terms of

binding energies, targeting the pY705 binding site at numerous receptor conformations, as well as decent agreement between the two softwares employed in ligand binding-poses predictions.

(III) Binding free energies of the STAT3:STAT3 and STAT3 dimer:DNA intermolecular association calculated by means of the MM/PB(GB)SA methods have shown to be corresponding for both pSTAT3-DNA and uSTAT3-DNA complex in terms of the protein-DNA interaction. However, the binding free energy of the protein-protein association is two-fold more favourable for the pSTAT3 complex than for the uSTAT3 complex. These predicted free energies of binding are in good qualitative agreement with experimental data, which demonstrated uSTAT3 protein binding to *ds*DNA, forming a stable conformation (Nkansah *et al,* manuscript submitted for publication, 2012).

(IV) A shared 3D-pharmacophore model of the pY705 containing tetrapeptide (P704-pY705-L706-K707) based on five representations of the pY-binding pocket (obtained from a 50-ns MD trajectory), confirmed the significance of residues pY705 and L706 in STAT3 binding, and their respective binding pockets. Collected specific features of the tetrapeptide - SH2 domain interaction/recognition may then provide a useful “guidance” in rational design of potential small molecule inhibitors of the STAT3-STAT3 interaction. Sterical restrictions of the pY-binding pocket (i.e the explicit protein residues) were also taken in account.

Supplementary information for CHAPTER 5:

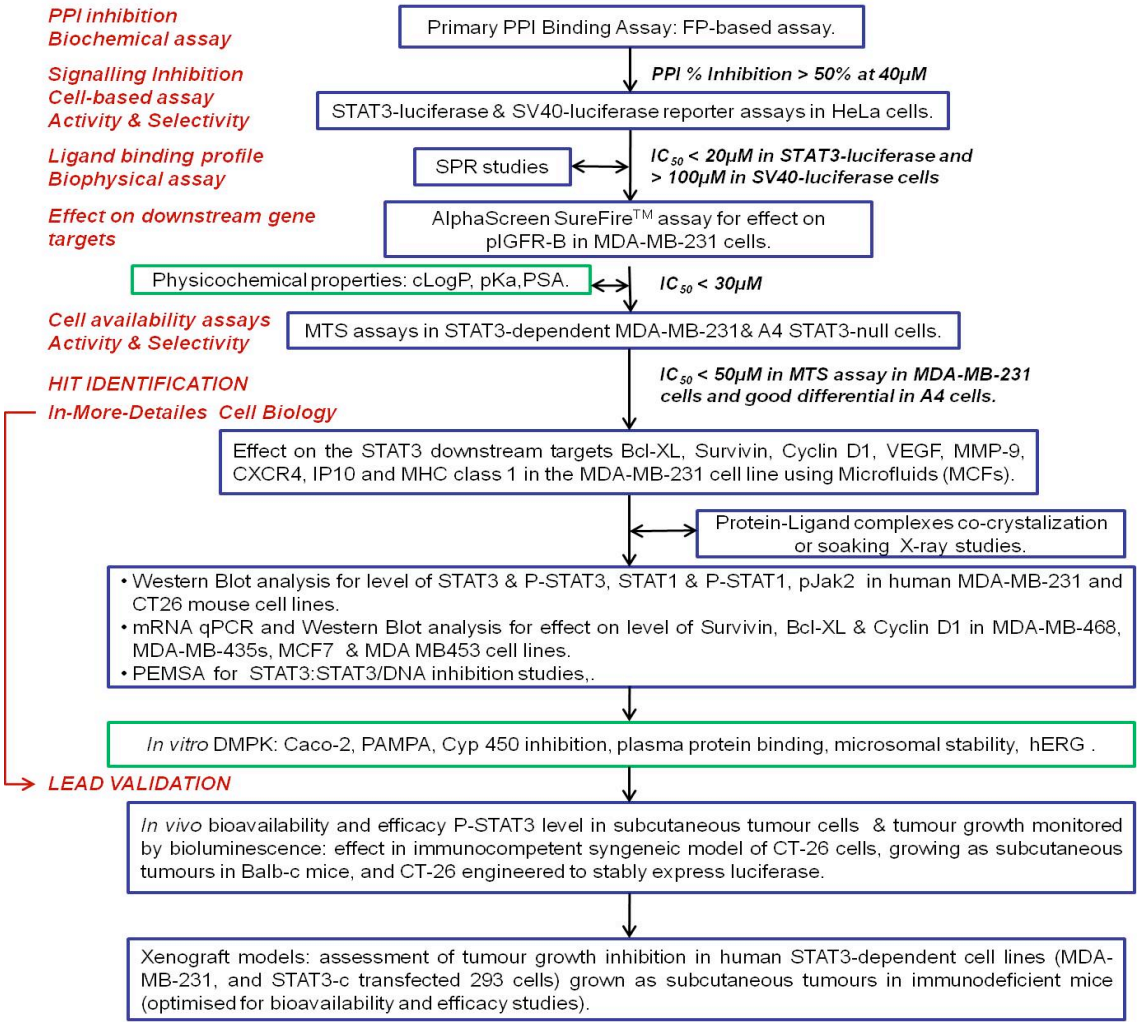


Figure S5.1: Overview of the full screening cascade for the potential STAT3 small-molecule inhibitor[8] The initial HTS with ~25.000 molecules was carried out at the European Screening Port (ESP) in Hamburg, employing a FP-biochemical cell-free assay as a primary PPI binding assay. Out of the 223 compounds selected by HTS (based on PPI % inhibition > 50% at 40 μM concentration), 54 compounds were purchasable, and brought into further biological experiments and *in silico* studies.

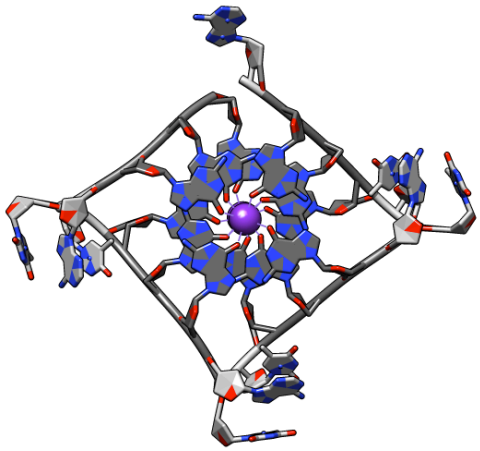
[8] The screening cascade overview was kindly provided by Dr. Giovanna Zinzalla

Table S5.1: Overview of the selected parameters employed in the DOCK6 docking protocol.

Molecule Library Input Parameters	
ligand_atom_file	/path/docking/ligand*.mol2
calculate_rmsd	no
Orient Ligand Parameters	
orient_ligand	yes
automated_matching	yes
receptor_site_file	/path/docking/selected_spheres.sph
max_orientations	500
Internal Energy Parameters	
use_internal_energy	yes
internal_energy_rep_exp	12
Flexible Ligand Parameters	
flexible_ligand	yes
min_anchor_size	5
pruning_use_clustering	yes
pruning_max_orients	100
pruning_clustering_cutoff	50
pruning_conformer_score_cutoff	25
use_clash_overlap	no
Bump Filter Parameters	
bump_filter	yes
bump_grid_prefix	/path/docking/grid
max_bumps_anchor	6
max_bumps_growth	6
Master Score Parameters	
score_molecules	yes
Grid Score Parameters	
grid_score_primary	yes
grid_score_grid_prefix	/path/docking/grid
Hawkins GB/SA Score Parameters	
gbsa_hawkins_score_secondary	yes
gbsa_hawkins_score_rec_filename	/path/docking/receptor-charged*.mol2
gbsa_hawkins_score_solvent_dielectric	78.5
gbsa_hawkins_use_salt_screen	yes
gbsa_hawkins_score_salt_conc(M)	0.15
gbsa_hawkins_score_gb_offset	0.09
gbsa_hawkins_score_cont_vdw_and_es	yes
gbsa_hawkins_score_vdw_att_exp	6
gbsa_hawkins_score_vdw_rep_exp	12
Simplex Minimization Parameters	
minimize_ligand	yes
minimize_anchor	yes
minimize_flexible_growth	yes
simplex_anchor_max_iterations	500
simplex_grow_max_iterations	500
simplex_secondary_minimize_pose	yes
simplex_secondary_max_iterations	100
Molecule Library Output Parameters	
ligand_outfile_prefix	ligand*
num_primary_scored_conformers_rescored	5
num_secondary_scored_conformers	3
write_secondary_conformations	yes
rank_secondary_ligands	no

PART 2

MOLECULAR MODELING STUDIES OF G-QUADRUPLEX DNA COMPLEXES WITH SMALL-MOLECULE LIGANDS



"The purpose of a model is not to fit the data but to sharpen the questions."
(Samuel Karlin, 1983)

'Overture'

Nucleic acid sequences containing multiple short sequences of guanosine nucleotides (G-tracts) can form higher-order structures, termed G-quadruplexes. These structures typically have a core of 2-4 G-quartets, each of which comprises four in-plane guanine bases in a highly stable hydrogen-bonded arrangement.[242,16] Successive G-quartets are held together by π-π stacking interactions, together with an array of metal ions coordinated to the O6 guanine substituents at the center of each quartet.[243] Quadruplex sequences occur at the telomeric ends of eukaryotic chromosomes. Thus in human telomeres a well-studied intramolecular quadruplex can be formed from four tandem repeats of the sequence d(TTAGGG). [242,244,14] Quadruplex sequences are also prevalent in other genomic locations,[73,245] notably in oncogenic promoters. In general the sequences connecting the short G-tracts may adopt structural linker roles, connecting the G-quartets in a particular manner. They can form loops (for example TTA loops in human telomeric quadruplexes), with relationships to the stack of G-quartets that depend on the overall topology of the quadruplex.[16] Structural information on G-quadruplex architecture has become available from X-ray crystallography, and from NMR structure determinations for human and a few other telomeric quadruplexes, and for a small number of promoter quadruplexes, for example for the c-MYC, c-KIT and b-RAF oncogenes.[76,246] These studies have revealed a high degree of structural diversity, with dependency on not only primary sequence but also on a number of environmental factors, notably the nature of the metal ion and quadruplex concentration.

The potential occurrence of quadruplex structures in cellular environments has suggested that they can be targets for therapeutic intervention, for example to down-regulate the transcription of cancer-associated genes or to inhibit the maintenance of telomeres in cancer cells.[74-76] A large number of small-molecule compounds have been devised and evaluated for quadruplex binding and as potential effectors of quadruplex cellular action. The majority of these compound classes have common structural features – planarity and cationic substituents, and attempts at developing more diverse structural types have in large part used library screening approaches.[99,247,248] Structural-based methods the rational design of more potent and selective compounds

are still in their infancy and have been hampered by the relatively small number of experimental structures available to date. *In silico* screening of libraries has been reported in several instances.

It is now widely recognized that many G-quadruplexes can adopt multiple conformations and topologies, and so a major challenge to the design of high-affinity ligands is the specificity of targeting G-quadruplexes with different folding patterns. It is known that particular small molecules can stabilize one topological type over another. For example, the anti-parallel hybrid topologies found in dilute solution for some intramolecular telomeric quadruplexes are stabilized by ligands such as the natural product telomestatin,[249] whereas the parallel topology can be stabilized by, for example, various tetra-substituted naphthalene diimides.[79] The majority of small-molecule compounds reported to date bind to the more general features of quadruplexes – the planar surface of a terminal G-quartet and the anionic phosphate groups, and the overwhelming majority bind to the planar G-quartet termini rather than predominantly in the grooves. Extended substituents are presumed to bind in the grooves and loop regions, although the available structural data on these features is limited to date. The discrimination between G-quadruplexes and other forms of DNA, in particular duplex DNA, is also of practical therapeutic importance, and is more straightforward to achieve. A fundamental goal remains the ability to design a small molecule that is highly selective for a particular quadruplex. Analysis of crystallographic data has shown that even for a given quadruplex type, the loops have conformationally flexible conformations, with bound small molecules inducing ligand-specific conformational change.[250] Thus loop flexibility adds another layer of complexity to the design challenge.

CHAPTER 6:
A multiple molecular dynamics approach for systematically examining conformational space in small-molecule/G-quadruplex interactions

6.1 BACKGROUND:

For rational drug design and effective G-quadruplex ligand optimization, it is of key importance to understand the interactions between small-molecule ligands and their G-quadruplex targets. X-ray diffraction and NMR experiments provide detailed insight into the structures of G-quadruplexs, and ligand-quadruplex complexes. However, they give little insight into the dynamic conformational rearrangements that small molecule–quadruplex complexes undergo, and the ligand transitions leading to G-quadruplex stabilization. Obtaining the most favourable binding poses of the compounds *in silico* has been challenging due to specific features of quadruplex nucleic acid molecules (i.e highly charged backbone, presence of stabilizing alkali metal cations, the basic quadruplex architecture) and in particular flexibility of the G-quadruplex structures is an important issue to be considered within the framework of G-quadruplex ligands docking. Quality of the G-quadruplex docking results, and subsequently the binding affinity of potential G-quadruplex ligands, might be strongly affected by the flexibility of the loop regions.[98]

The present study has approached the problem of probing and predicting the low-energy arrangements and conformational profile for small-molecule-quadruplex binding by developing a set of computational tools for examining large ensembles of potential binding sites and loop conformations. The intramolecular human telomeric quadruplex structure has been used as the quadruplex platform, since the native structure and complexes with a wide range of ligands have been determined by X-ray crystallography. The methodology though is applicable to all quadruplexes for which even a low-resolution topological description is available.

Two contrasting small molecule ligands have been used as probes. The compound pyridostatin,[251] based on a N',N''-bis(quinolinyl)pyridine-2,6-dicarboxamide scaffold, has high selectivity for quadruplex vs duplex DNA. Its ability to target telomeric quadruplexes has been demonstrated[252] and it has also been show that pyridostatin interacts selectively with a terminal G-quartet of a G-quadruplex through a stacking mode. Furthermore, Rodriguez[252] *et al* have also provided evidence that pyridostatin targets the common structural feature shared by G-quadruplex motifs regardless the nature of the loop sequences. This ligand has itself a high degree of conformational flexibility, with a total of 13 rotatable bonds. Although it has been suggested that the pyridostatin molecule would have a preferred near-planar conformation as a result of potential coordination of four inner-facing nitrogen atoms to a water molecule, the present study has not assumed that this would occur. By contrast the pentacyclic acridine derivative RHPS4[253] has no rotatable bonds and just methyl and fluorine substituents. RHPS4 and a number of its analogues bind to human telomeric quadruplex DNA with high affinity and some selectivity for quadruplex compared to duplex DNA, although the measure of selectivity is dependent on the methodology used. An NMR study of RHPS4 bound to the tetramolecular quadruplex $d(TTAGGGT)_4$ has shown that all four strands are parallel in this complex[254] and has provided an experimental structure to which simulation studies can be compared.

6.2 AIMS:

The aim of this work is to develop a systematic and general approach to determining all plausible positions for binding a ligand to a G-quadruplex structure, taking fully into account the flexibility of both the target and the ligand, as well as any conformational change upon ligand binding, and then evaluating the energetically most favourable binding regions of the resulting complexes using multiple free-energy of binding calculation techniques.

6.3 METHODS:

The four principal stages in the approach involve (1) in silico construction and preparation of the two ligands, (2) sampling of the G-quadruplex/ligand conformational space via calculations of explicit MD full-atomistic trajectories of the G-quadruplex/ligand complexes starting from numerous ligand positions, (3) clustering of the ligand conformations determined by the MD calculations, and (4) calculation of binding energies of the proposed binding poses of the complexes using two different methods.

6.3.1 System setup and molecular dynamics simulation

Molecular structures of pyridostatin[251] and RHPS4[254] were constructed using the ChemBioOffice (Cambridgesoft) package, and saved as a Sybyl 'mol2' file format. Since RHPS4 is a rigid molecule (Figure 6.1 a) there were no intramolecular conformations to explore. However pyridostatin has a number of flexible bonds (13 rotatable bonds in total) which allow the molecule to adopt a wide range of different conformations. For the starting positions it was assumed that the central part of the pyridostatin molecule was planar (Figure 6.1 b). There were several possible conformations where the aromatic systems and the peptide linkages could be co-planar. Each peptide linkage has three rotatable bonds, and since we are assuming that the molecule is flat, each of the central rotatable bonds has two possible conformations. Because there are six of them in total, there could be 2^6 (or 64) conformations. However for steric reasons, only 16 conformations were possible (Table 6.1 in the results section).

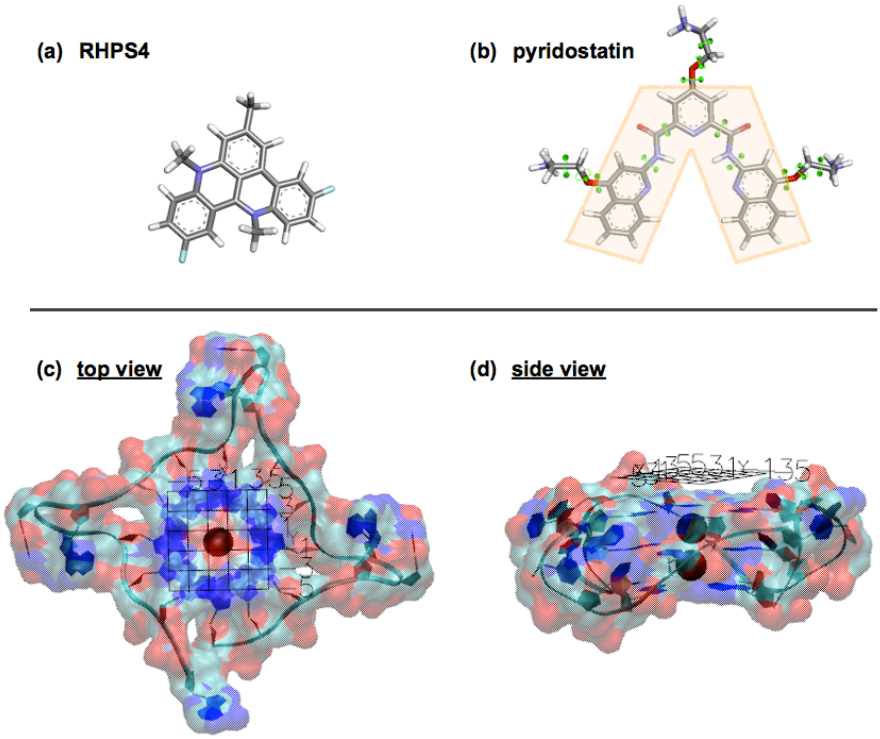


Figure 6.1: Models of the ligands, and the human telomeric G-quadruplex structure (22-mer).
(a) rigid RHPS4 molecule and (b) flexible pyridostatin molecule. The central "flattened" atoms of pyridostatin in one of its possible conformations, highlighted in the yellow panel; (c, d) human telomeric G-quadruplex in surface representation with the 2 Å grid aligned with the least squares plane of the guanine bases; The grid is placed 3.4 Å above the plane (2 Å above the surface), 5'site is shown here. These grid points on both 3'- and 5'site were used as starting points for multiple MD-runs.

The crystal structure of telomeric G-quadruplex[15] (PDB id 1KF1 at 2.1Å resolution) was used as a starting-point for the molecular dynamics studies. Consecutive K+ ions vertically aligned within the central core of the G-quadruplex mid-way between each G-quartet were retained at their respective crystallographic positions, while the K+ ion outside of the central core of the G-tetrad (at the 5'-site) was removed. The positions of each ligand were explored on both the 3'- and 5'-face of the quadruplex molecule. The G-quartet on the binding site of the quadruplex was aligned with the x,y axes by firstly calculating the least-squares plane of the guanine base atoms. Eigenvalues for the least-squares plane calculation were obtained using the linear algebra module from the NumPy[255] package. The G-quartet was then oriented accordingly. The position of the center of mass of the atoms belonging to bases of the G-quartet on the binding site of the quadruplex was calculated, and the molecule translated so that this lay on the origin. Each of the 16 planar conformations of the (flattened) pyridostatin molecule was aligned to lie on the same plane as the x-y axes by calculating the least-squares plane of

the central atoms and rotation accordingly. The molecules was then translated so that the center of mass of the planar central scaffold was on the origin and they were translated in the z-direction so that they lay 3.4 Å above the least-squares plane of the chosen quadruplex binding site. These positions were uses as a reference for any further translations and rotations (Figure 6.1 c,d). The single structure of the RHPS4 molecule was prepared in the same way. Each molecule was then rotated in 45° steps to produce eight orientations and each oriented molecule was translated to positions on the x-y plane in a grid-like manner to produce 36 different positions at 2 Å intervals in both the x and y directions (-5, -3, -1, 1, 3, and 5 Å form the original position in both the x and y direction). Each position was then tested for clashes with the quadruplex and all those with any ligand-quadruplex separations of less than 2.4 Å were rejected.

All of the MD simulations were full-atom simulations and were performed with the GROMACS v 4.5.3 program,[154,155] employing the *parmbsc0*[161] force field, which was ported previously into GROMACS (as described in CHAPTER 3). The topologies and other parameters for the two ligands were generated using the ACPYPE[169] tool, which employs the ANTECHAMER[170] module of the AMBER11 program with the Generalized Amber Force Field (GAFF).[226] The simulation protocols were consistent for both 22-mer quadruplex ligand-bound systems. The models were solvated in a periodic box of TIP3P[196] water molecules, with a minimal clearance of 20.0 Å between periodic images for the starting configurations. Additionally, positively charged K+ counter ions were included in the systems to neutralize the negative net charge on the DNA backbone. Subsequently the systems were subjected to 10,000 steps of robust potential energy minimization (steepest descent followed by conjugate gradient algorithm), followed by 100 ps of molecular dynamics at 200K while keeping the solutes constrained. The systems were then slowly heated to 300K, with further unconstrained equilibration steps over 50 ps. Production-level MD trajectory calculations were carried out at 300K with a time constant for coupling of 0.1 ps under the control of velocity rescaling thermostat,[199] and isotropic constant-pressure boundary conditions controlled by the Parinello-Rahman[200,201] algorithm with pressure coupling at 1.0 bar. Non-bonded van der Waals interactions used the Lennard-Jones 12-6 potential with a 10.0 Å cut-off, and the particle mesh Ewald (PME)[122] method was employed for electrostatic interac-

tions. All-atom bonds were constrained by the LINCS[126] algorithm. The time step applied was 2.0 fs, with coordinates saved every 5.0 ps. All the MD simulations were run in 100 ps blocks, up to 1 ns, which was considered to be an equilibrium position.

6.3.2 Cluster analysis

After each block of the 100-ps explicit solvent MD simulations, an average structure over the final 10 ps was generated. This structure was then aligned against all of the preceding averaged structures for that starting conformation. Two opposite guanine bases on the G-quartet adjacent to the bound ligand molecule were used as a reference for the alignment (the guanine bases of the quadruplexes were found to be relatively stable during the simulations). The bases were aligned by calculating the principal moments of inertia and the molecules were then translated so that the center of mass of the chosen guanine bases was at the origin and the molecules were rotated so that the principal axes of inertial corresponded to the coordinate axes. The structures were then clustered, based on the positions of the central part of the ligand molecules. The root-mean-square distance (RMSD) of the central atoms between each pair of averaged structures was calculated. Hierarchical agglomerative clustering was then carried out using the nearest linkage method to an RMSD cutoff of 1.2 Å. If a ligand was found to share a cluster with a previous average structure or had been run for a total of 1 ns, a simulation was begun with a new starting position. Eventually, all the truncated and completed MD runs were clustered based on the criteria described above.

6.3.3 Binding energy calculations

Two different methods were used to calculate the free-energy change on quadruplex-ligand complex formation, for all final conformations sampled throughout the numerous MD simulations (dynamic docking): (1) using the semiempirical quantum chemistry program MOPAC2009 v.11.3 (http://openmopac.net) with the MOZYME[256] linear-scaling algorithm (2) using the force field-based molecular mechanics/Poisson-Boltzmann (or generalized Born) surface area (MM/PB(GB)SA)[131,132,229] methods as implemented in MMPBSA.py python module within the AMBER11 package. Binding energies of a small subset of the structures were calculated employing the force field-based Generalized Born/volume integral (MM/GBVI)[257] method within the MOE package (http://www.chemcomp.com) were also used as an additional, third method, for a comparison purposes.[9] All binding energy calculations were performed using the same set of structures, in each case five frames representing the last 10 ps of the final 100-ps block of the individual simulations. Free-energy changes of the energy-minimized systems were also calculated with MOPAC2009, to evaluate the initial G4/ligand interactions prior to MD simulations. Binding energy calculations by means of MOPAC2009 calculations were employed in order to evaluate the correlation/differences between semi-empirical and empirical methods (MM/PB(GB)SA), and to obtain statistically stronger set of calculations for the subsequent analysis.

• the PM6-DH2 method[258] (with correct dispersion and hydrogen bond terms), together with a single-point calculation (1SCF), was employed within MOPAC2009 to calculate the interaction energy between the quadruplex and the ligands. The effect of a solvent model surrounding the molecules was approximated via the COSMO method[259] with a dielectric constant for the implicit solvent set to 78. The intermolecular interaction energy E_{INTER} was then given by calculating the heats of formation of each of the three systems; G-quadruplex/ligand complex (G4-LIG), G-quadruplex (G4) and the ligand (LIG) (equation 6.1)

[9] I acknowledge Dr. J.A. Platts for the valuable advice and mentoring in free binding energy calculations by different approaches described in this chapter.

$$E_{INTER} = \Delta H_{f(G4\text{-}LIG\ complex)} - \Delta H_{f(G4)} - \Delta H_{f(LIG)} \quad (6.1)$$

where $\Delta H_{f(G4\text{-}LIG\ complex)}$, $\Delta H_{f(G4)}$ and $\Delta H_{f(LIG)}$ are the heat of formation of the complex, G-quadruplex and ligand respectively. Five frames of the MD trajectory, representing the final 10 ps of the final 100-ps 'blocks' of each of the conformations, positions and orientations were saved as pdb files and subsequently used as input for the MOPAC binding-energy calculations. A single value of the calculated intermolecular interaction energy was obtained as an average over the five frames.

• the MM/PB(GB)SA[131,132,229] method computes the relative free energies of binding, employing the thermodynamic cycle that combines the molecular mechanical (MM) energies with the implicit solvent methods. This method takes advantage of multiple snapshots from a trajectory, to provide an average of energies. The change of free-energy of the molecules upon complex formation was calculated (for each of the snap-shots) as a difference of free energy between their bound and unbound states is given in equation 6.2 (more details of the MM/PB(GB)SA method are provided in CHAPTER2); free energy of each term was then calculated according to equation 6.3:

$$\Delta G_{bind} = G_{(G4\text{-}LIG\ complex)} - G_{(G4)} - G_{(LIG)} \quad (6.2)$$

$$\Delta G = \Delta E_{MM} + \Delta G_{SOL} - T\Delta S \quad (6.3)$$

where the molecular mechanics energy (E_{MM}) term is a sum of the internal energy (bonds, angles and dihedrals), electrostatic energy and van der Waals term; while the G_{SOL} term accounts for the solvation energy, comprising both polar and nonpolar component. The polar part of the solvation term accounts for the electrostatic contribution to solvation and was calculated using both Poisson-Boltzmann (PB) model, and the generalized-Born (GB) model developed by Onufriev *et al*[234] (igb=2; model GB^{OBC1}) with rescaled effective Born radii, accounting for interstitional spaces between atom spheres. Solvent probe radius, dielectric constants grid-spacing parameters were kept at program's default values. The entropy term (TΔS) was not included in our simulations as different conformations/ binding poses of one ligand were explored. All the calculations were performed employing the single-trajectory approach, and used the five frames of the final 10 ps (corresponding to the approach for the MOPAC calculations).

Prior to the binding free energy calculations, the individual trajectories (stripped from the explicit solvent) were converted from GROMACS compressed trajectory 'xtc' file into AMBER trajectory file format ('mdcrd' file) via VMD visualization program, and via AMBER trajectory processing program *ptraj*.

• the generalized Born/volume integral (GB/VI)[257] method that estimates the free energy of hydration (together with cavitation energy), based on a VI London dispersion energy (unlike the GB/SA hydration model, that is based on atomic surface area, SA) employed with the MOE (an acronym for Molecular Orbital Environment) software package was used as the third method to estimate the free energy of binding of the G4/ligand complexes. This method was applied only to a limited number of G4/ligand conformations, serving rather as a 'test method'.

6.3.4 Statistical analysis

The mean binding energies for each cluster were calculated. In order to determine the statistical significant of the differences between these, one way analysis of variance (ANOVA) was carried out using the statistical package R,[260] considering all clusters that had three or more members. RMSD calculations for the guanine bases proximal to the ligand were calculated using the program LSQMAN.[209] The variations within clusters were examined for each cluster with three or more members, by determining the RMSD between each pair of cluster members and calculating the mean RMSD value. The variations between clusters were also calculated and each cluster was compared with every other cluster. Comparison of two clusters was performed by calculating the RMSDs of each member of the first cluster (cluster A) with each member of the cluster that it is being compared with (cluster B), and the cluster similarity was taken to be the mean RMSD. So for example if cluster A had 8 members and cluster B had 3 members, a total of 24 RMSD calculations were carried out and the mean value of these RMSDs was considered to be the similarity of the two clusters.

6.4 RESULTS AND DISCUSSION:

The aim of this work is to develop a systematic and general approach to determining all plausible positions for binding a ligand to a G-quadruplex structure, taking fully into account the flexibility of both the target and the ligand, as well as any conformational change upon ligand binding, and then evaluating the energetically most favourable binding regions of the resulting complexes using multiple free-energy calculation techniques.

Several levels of analysis are presented here:

(1) two structurally very different ligands, the rigid RHPS4 molecule, and the conformationally flexible pyridostatin molecule, were examined for all sterically-plausible conformations, as well as for orientational and positional flexibility with respect to the G-quartet faces of the quadruplex.

(2) both 5' and 3' termini of the human telomeric quadruplex were explored in order to avoid any false-positive binding on one of the sites, as well as to inspect whether there is an energetic preference for either of the binding sites.

(3) the changes in free energy resulting from intermolecular interaction of the quadruplex with ligand were calculated, using a semiempirical quantum chemistry level technique (MOPAC), and molecular mechanics (force field-based) methods (MM/PB(GB)SA and MM-GB/VI). The semiempirical (MOPAC) calculations were performed for a comparison purposes, complementary to the empirical MM/PB(GB)SA approach. Subsequently, the correlations between the methods were statistically and graphically examined.

(4) the structural-stability effects of the ligands on the G-quadruplex structure were inspected by means of RMSD calculations of the non-hydrogen atoms within the 3'/5' termini G-quartets. The structural differences within the clusters, as well as in between the clusters, were calculated.

The overall workflow of the MD simulations ('dynamic docking'), structural alignment and clustering, subsequent free energy calculations on quadruplex-ligand complex formation, and the statistical and structural analyses, is represented in Figure 6.2.

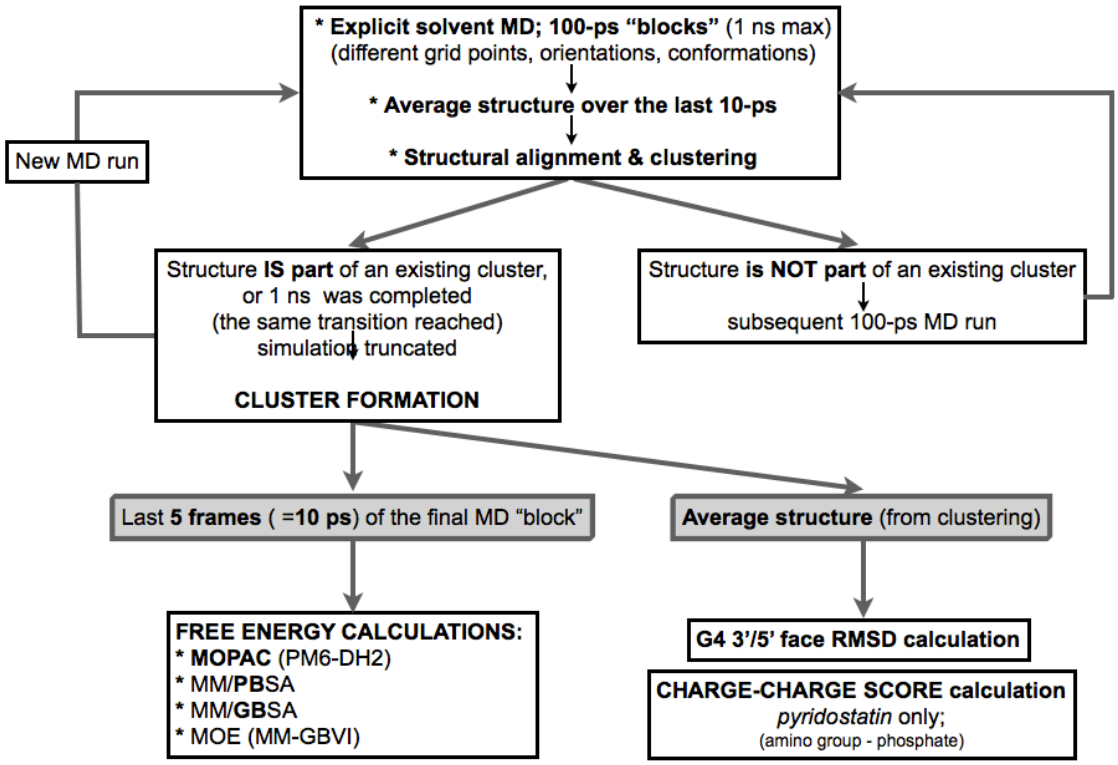


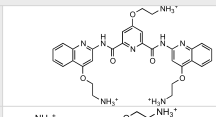
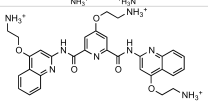
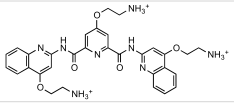
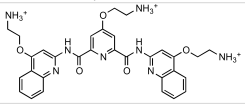
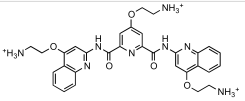
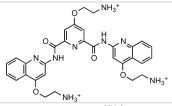
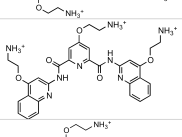
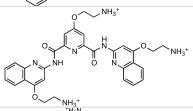
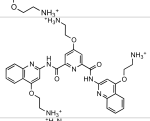
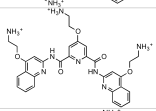
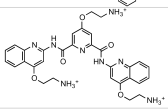
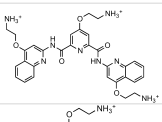
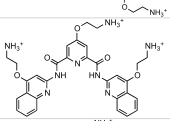
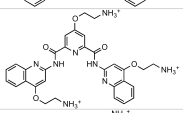
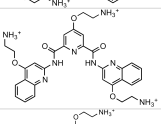
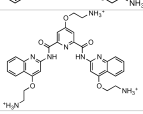
Figure 6.2: Schematic overview of the workflow.
Upon numerous MD simulations employing different ligand's conformations, orientations and positions on the G4 3' and 5'face, time-averaged structures over the final 10 ps of the last 100-ps blocks used for clustering were further employed for the target's face stability calculation (heave atoms only), as well as for the charge-charge score calculation between the positively-charged side chains of the ligand and the negatively-charged G4-DNA backbone; five frames representing the final 10 ps of the 100-ps MD runs were then used for the MM/PB(GB)SA calculations, as well as for the MOPAC calculations.

The individual components of the procedure were automated via in-house python scripts[10], in order to cope with the very large number of starting structures (2436 in total) and extensive output data. Explicit-solvent molecular dynamics simulations were performed, with 232 starting positions for the quadruplex-RHPS4 complexes, and 2204 individual starting positions of pyridostatin in 16 different conformations (Table 6.1) over both 3' and 5' ends of the quadruplex (Figure 6.1). The disproportion in the amount

[10] The automation of the simulation procedure via python scripts was performed by Dr. Alan K. Todd

of MD simulations started for the two ligands is due to the nature of their structures; since pyridostatin is highly flexible and conformationally variable, the amount of possible binding modes exceeded the number of RHPS4's positions nearly 10-fold. Furthermore, the difference in the overall number of simulations performed on the 3' and 5'-site respectively is possibly due to the difference in shape of the 3’- and 5’site; the latter being significantly flatter, thus providing larger surface to sample. Table 6.1 further summarizes the trend of the individual pyridostatin conformational transitions throughout the MD runs.

Table 6.1: Overview of the 16 explored pyridostatin conformations.
These 16 conformations of pyridostatin were examined by an extensive MD study (i.e fully dynamic docking) for both 3'- and 5'face of human telomeric G-quadruplex. A number of the started MD runs, a number of 100-ps blocks performed for each of the conformations, and the percentage of gain/loss of that particular conformation throughout the multiple MD runs is given.

Pyridistatin conformations explored		5'site MD simulations			3'site MD simulations		
		# MD runs started	# 100-ps blocks	Conform. gain/loss	# MD runs started	# 100-ps blocks	Conform. gain/loss
000000		105	536	+ 32%	17	123	+ 165%
000001		104	561	+ 14%	26	200	− 15%
000010		116	570	− 3%	40	287	− 12%
000011		90	595	− 50%	38	256	− 50%
000100		121	565	+ 21%	25	225	+ 64%
000101		99	476	+ 15%	18	145	+ 39%
000110		101	594	− 8%	38	221	− 13%
000111		116	586	− 28%	34	237	− 53%
001000		117	474	+ 11%	26	214	+ 50%
001001		104	507	− 9%	43	283	− 14%
001010		111	512	+ 2%	8	64	+ 75%
001011		112	509	− 6%	30	229	− 43%
001100		119	433	+ 16%	33	154	+ 12%
001101		120	535	+ 7%	18	129	+ 28%
001110		110	414	0	28	189	− 4%
001111		124	543	− 22%	13	119	− 62%

With 32% (5'site) and 165% (3'site) conformational "gain" throughout the MD runs, conformation '000000' was shown as favourable by both G-quadruplex faces, with the largest increase of this conformation by a conversion of other, less favourable, pyridostatin conformations explored. Similarly, conformations '000100', '000101', '001100', '001101' and '001010' have increased in their numbers throughout the simulation time. On the other hand, conformation '000011' showed the largest loss of structures (~50% on both 3' and 5' sites) throughout the MD runs, converging to other preferable conformations. Table 6.2 provides an overview of the total number of simulations started, completed (i.e reached 1 ns), and truncated throughout the MD run for both ligands on each site of the G4. In terms of the rigid ligand, RHPS4, less than 10% of all the simulations that started, reached the maximum simulation time of 1 ns, and over 90% were even truncated prior to/at 0.5 ns. This situation applies to both 5' and 3' face of the G-quadruplex, and may be understood as a consequence of the rigidity of the RHPS4 compound, together with the presence of no rotatable bonds, resulting only in a little conformational requirements of the ligand. A different trend is observed for the conformationally flexible ligand, pyridostatin; on the 5'-face of the G-quadruplex, over 60% of the MD runs were truncated, as a result of clustering, halfway through the maximum length of 1 ns, and only ~ 30% completed the 1-ns run. The trend on the 3' site is reversed, with ~60% of the MD runs completing the full 1-ns trajectory, while only about 30% runs are truncated prior to/at 0.5 ns. This finding may be a consequence of the 3'site shape (i.e more "bowl"-like shape), as well as the conformational demand of pyridostatin molecule.

Table 6.2: Overview of the results and statistics for the two studied G4/ligand complexes.
The rigid RHPS4 was compared with the highly flexible pyridostatin. Both 3'- and 5' sites of the telomeric G-quadruplex structure (22-mer, PDB id 1KF1) were examined.
(* p-value is a statistical value that details how much evidence there is to reject the most common explanation for the data set)

	RHPS4		**PYRIDOSTATIN**	
	5'site	**3'site**	**5'site**	**3'site**
No simulations started	141	91	1769	435
No 100-ps blocks	275	207	8184	3075
Completed 1 ns	2%	10%	31%	59%
Truncated prior 0.5 ns	94%	90%	64%	35%
MOPAC mean [kcal/mol]	-24.76	-26.37	-39.50	-36.77
MM-GBSA mean [kcal/mol]	-27.04	-27.35	-43.25	-39.53
MM-PBSA mean [kcal/mol]	-25.75	-26.36	-39.49	-37.33
No clusters ≥3	20	7	138	26
Structures within clusters	105 (75%)	73 (80%)	961 (54%)	119 (27%)
MOPAC ANOVA grand mean [kcal/mol]	**-24.87**	**-26.87**	**-40.74**	**-36.61**
MOPAC ANOVA P-value	1.38×10^{-11}	0.010	2.2×10^{-16}	1.36×10^{-5}
MM-GBSA ANOVA grand mean [kcal/mol]	**-27.08**	**-27.52**	**-44.28**	**-40.15**
MM-GBSA ANOVA P-value *	0.022	0.269	2.2×10^{-16}	2.13×10^{-4}
MM-PBSA ANOVA grand mean [kcal/mol]	**-25.75**	**-26.43**	**-40.49**	**-37.30**
MM-PBSA ANOVA P-value	0.053	0.444	2.2×10^{-16}	9.14×10^{-7}
Correlation MOPAC - MMGBSA	0.46	0.74	0.835	0.825
Correlation MOPAC - MMPBSA	0.51	0.52	0.882	0.917
Correlation MMGBSA - MMPBSA	0.70	0.72	0.876	0.860

Graphical representation of the MD runs in their 100-ps "blocks" and their truncation along the simulation time, hence the clustering network, is shown in Figure 6.3 for one of the pyridostatin conformations (conformation termed as '000000') explored.

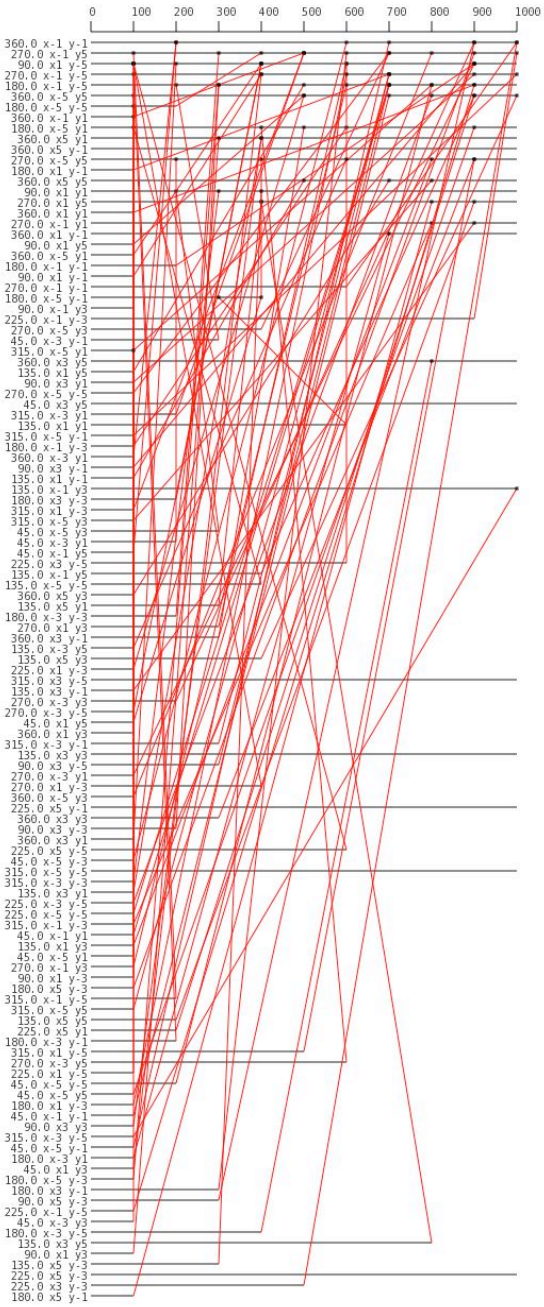


Figure 6.3: Diagram showing truncated completed MD-runs of one of pyridostatin's 16 conformations. Conformation termed as '000000' on the 5'site is shown; The initial grid positions and rotations of the individual ligand structures are indicated. The length of corresponding lines represents the length of each of the simulations, 100 – 1000 ps, with the red lines representing the clustering network.

6.4.1 Clustering the multiple molecular dynamics simulations

The clustering of the MD runs was performed on the basis that once ligand has adopted a certain conformation, which was already represented in a pool of structures within other clusters, the subsequent ligand's transition would be alike, hence the MD run would be truncated, allowing for a new simulation to begin from a different starting point on the grid and in different orientation and/or conformation. Positioning the ligand along the grid with 2 Å spacing and 45 degrees rotations (orientations) allowed fine-sampling of the G-quadruplex surface with sterical requirements of the macromolecule being taking into account.

To statistically determine the similarity and difference between individual clusters of structures, one way analysis of variance (ANOVA), was carried out. In principle, ANOVA is a statistical method that allows simultaneous comparisons between two or more means, thus it determines whether or not the means of the studied variables are all equal. Here, only clusters that contained three and more members were considerd; there were 20 RHPS4-clusters on the 5'site and seven RHPS4-clusters formed of three or more members, which represented ~80% of all generated RHPS4-clusters. In the case of the pyridostatin-formed clusters, there were 138 clusters with three or more members on the 5'site, comprising ~55% of the structures generated, and 26 clusters ≥3, spanning only ~30% of the structures produced upon structural alignment and clustering. This observation suggests that both the shape of the sampled G4-site, as well as the conformational richness of pyridostatin plays a major role in adopting a particular binding pose of the ligand. However, mean binding energies obtained for these clusters via ANOVA analysis correspond closely to the mean values as calculated for all structures, without the consideration of the cluster size (Table 6.2) This will be further discussed in the following section.

The largest clusters of structures obtained from the MD runs are represented in Figure 6.4, including the transition paths from the ligand's starting positions on the grid, toward the newly-formed cluster of structures. The largest cluster of RHPS4 structures

placed on either site of the G-quadruplex (prior to MD simulations) comprised 10 structures on the 5' face, and 25 structures on the 3' face. Pyridostatin formed a 33-member comprising cluster on the 5' site, while the largest cluster on the 3' site was formed of 13 structures. The 5‘face of the G-quadruplex is flatter, comparing to the 3‘face that is more curved, thus it provided more “open space” where the ligands in their initial conformation were placed, avoiding the steric hindrance of the receptor. The molecules within a cluster are likely to follow a corresponding transition. To bring the positions of clusters and the orientation of the ligand within individual clusters (with three or more members) into visual perspective, simplified representations of the two structurally different ligands were created. These simplified representations with a “boomerang-like” resemblance are shown in Figure 6.5, and their triangular shape is defined by simple lines, connecting C05, C17 and C22 atoms of RHPS4 (Figure 6.5 a); and atoms C30 and C18 with the middle-point of the distance between atoms C4 and N4 of pyridostatin (Figure 6.5 b). All clusters with three and more structures were then visualized altogether on the grid, with each of the clusters represented by a single bead (the bead represented the centre of mass of the ligand). This provided an insight into the overall trend of clusters formation on the G-quadruplex 3' and 5' site, as shown in Figure 6.5. The bead representation however only served as one-dimensional description, without a sense for the ligand's orientation with respect to other clusters as well as the G-quadruplex. Using the boomerang-like representation of the ligands, whose boldness (thickness) reflected the size of the clusters, while preserving the information about the size of the ligand relative to the G-quadruplex (pyridostatin is a larger ligand, hence larger boomerang) the preferred orientation of the ligands was shown.

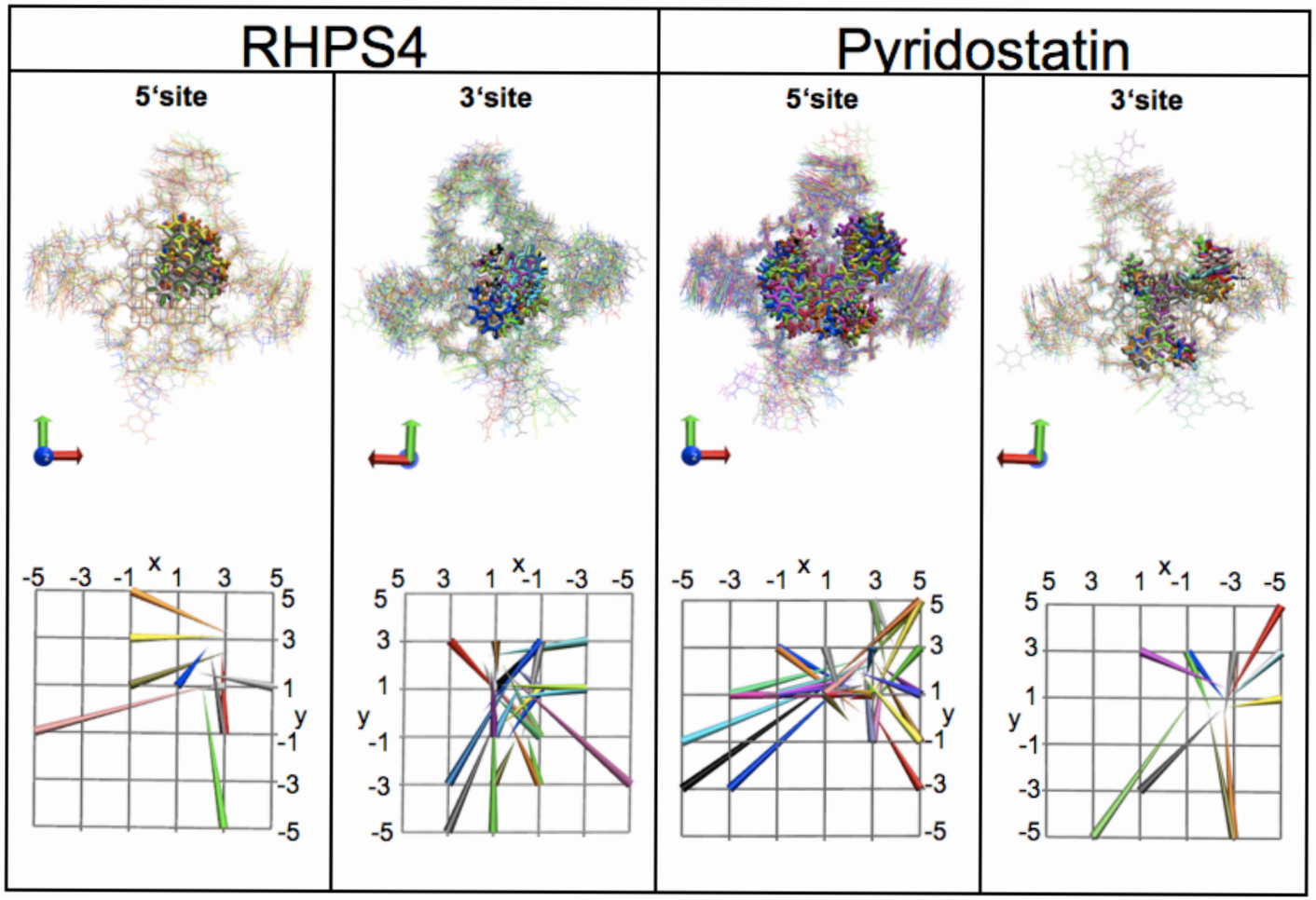


Figure 6.4: The largest clusters of structures obtained from the MD trajectories at the end of their MD runs (with the RMSD cutoff of 1.2 Å). The arrows on the grid are pointing from the positions where the ligands were placed prior the MD simulations and subsequent clustering.

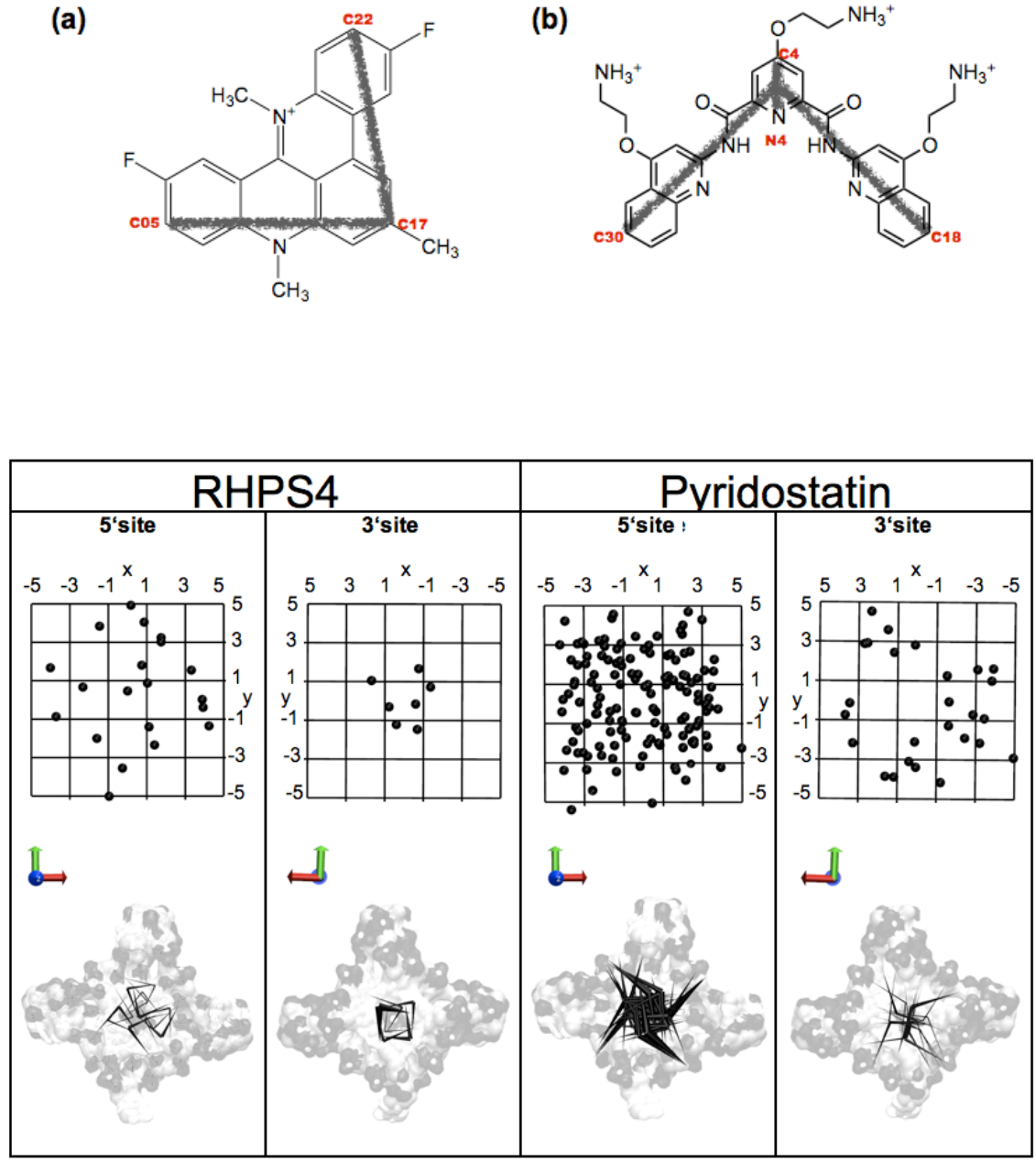


Figure 6.5: Clusters of structures of the of the G4/RHPS4 and pyridostatin/G4 complexes.
Simplified representation of the examined ligands, (a) the rigid RHPS4 and (b) the conformationally variable pyridostatin, are for the visualization purposes displayed as "boomerangs", created by connecting atoms C05, C17 and C22 atoms of the RHPS4 molecule; and atoms C30 and C18 with the middle point connecting atoms C4 and N4 of pyridostatin molecule. Clusters of structures (≥3) of the G4/RHPS4 and the pyridostatin/G4 complexes are shown for both 3'- and 5'site of the quadruplex, upon structural alignment and clustering. To reduce the dimensionality of the complex systems, each of the clusters is represented by a single "bead" per cluster located on the grid (top), and by a boomerang (triangular) representation, where the orientation of the resulting clusters is preserved/shown. The boomerang's size is proportional to the actual size of the ligand with respect to the G-quadruplex, and thickness of the "boomerangs" represents the size of the cluster.

6.4.2 Free energy calculations: semiempirical versus empirical methods

Intermolecular interaction energies of the G-quadruplex/ligand complexes were calculated for each of the final positions reached by the ligands in particular conformation, orientation and starting position on either site of the G-quadruplex. The means of the binding energies obtained by semiempirical approach (MOPAC) and molecular mechanics-based MM/PB(GB)SA methods as implemented in AMBER11, are summarized in Table 6.2, and graphically represented in Figures 6.6 and 6.7. To ensure consistency and comparability of the results, all calculations were performed using corresponding five frames representing the very last 10-ps of each completed simulation. The MM/PB(GB)SA calculations were performed over short trajectories (five frames, last 10 ps of each final simulation) upon the trajectory file format conversion, and MOPAC calculations were performed over five separate entries (for each final conformation), and their free energy values were averaged into one representative value.

Binding energies as calculated by means of MOPAC, and MM/PB(GB)SA, were plotted for each cluster of structures (≥ 3). The clusters of the flexible pyridostatin ligand are somewhat energetically different, with the extreme difference of ~25 kcal/mol between two clusters on the 5‘face (for instance a six-membered cluster #8 and a five-membered cluster #46) but no trend of more/less favourable energies with respect to a specific region on G-quadruplex were found. The calculated binding energies are in a good agreement among the methods for both of the ligands bound to either terminal site of G-quadruplex (Table 6.2), but there was a considerably higher level of correlation of the methods found for pyridostatin compared to RHPS4. This might be a consequence of (1) ligand’s flexibility and its conformational variability adopted by pyridostatin upon binding, and (2) charge-charge interactions (phosphates of the G-quadruplex with positive side chains of pyridostatin) that contribute to binding and stabilization of G-quadruplex. While pyridostatin might have a preference for the 5‘site of G-quadruplex, by ~4 kcal/mol when compared to the 3‘site, no significant difference in RHPS4 binding energies was found between the two sites of the G4, suggesting that there may not be a strong preference of the ligand for either of the sites.

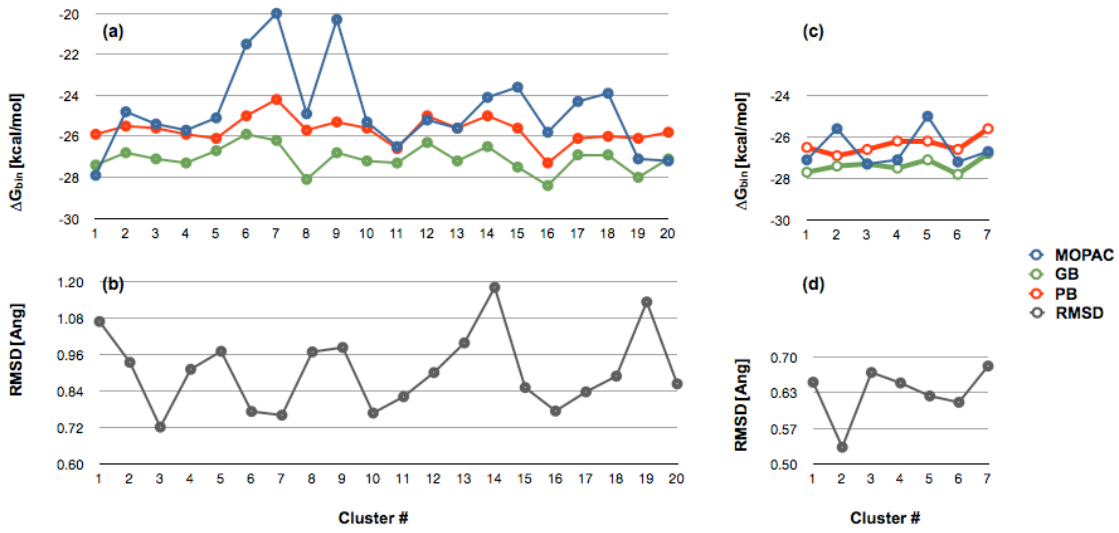


Figure 6.6: Calculated binding energies of the RHPS4 clusters with G-quadruplex structure.
Binding energies calculated by MOPAC (*blue*), MM/PBSA (*red*) and MM/GBSA (*green*) for RHPS4 clusters, obtained by multiple MD-runs are plotted for individual clusters of structures (≥ 3) upon structural alignment and clustering on 5' (a) and 3' site (c) of G-quadruplex. The dynamic response of the 3' (b) and 5' (d) face of the G-quadruplex is shown as mean RMSd value for each cluster (≥ 3), obtained by averaging all the RMSd's of structures within clusters (*grey*).

Since the side chains of pyridostatin carry a positive charge (the overall net charge +3), while the backbone of G-quadruplex has a negative charge, an independent method for evaluation of the charge-charge contribution to binding was performed by measuring the inverse square distances between each of the positively-charged side chains and the negatively-charged phosphates of the DNA backbone within 8 Å cutoff distance. The resulting score provided insight into the electrostatic contribution to the pyridostatin-G4 interaction. (Figure 6.7 a, d). Mean RMSD values within each of the clusters were also brought into perspective with the binding-energy and charge-charge contribution to binding (pyridostatin only) graphs (Figure 6.7 (b, d) and 8 (c, f). The rational behind stability calculation of the G-quartets on either 3' or 5' site was to quantify the response of G-quadruplex to the ligand binding with respect to the strength of the interaction. G4-stem is generally stable, while the loops are very flexible even in the native G-quadruplex structure, and so RMSD of the loops only would not realistically reflect the effect of ligand binding. Both ligands were found to have a stabilizing effect on G-quadruplex structure, when compared to G-quadruplex stability with no ligand bound.

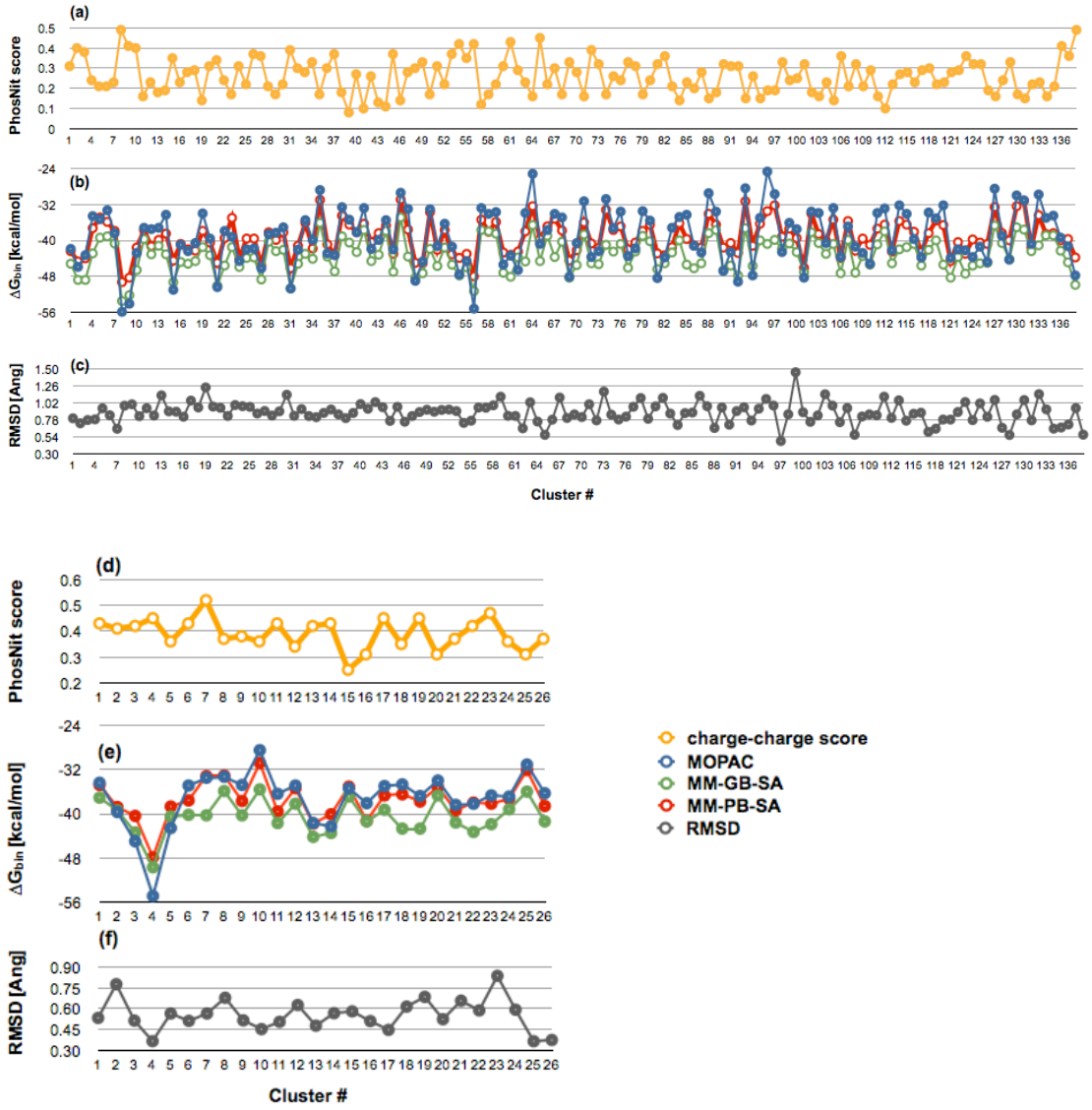


Figure 6.7: Calculated binding energies of the pyridostatin clusters with G-quadruplex structure.
Binding energies calculated by MOPAC (*blue*), MMPBSA (*red*) and MMGBSA (*green*) for pyridostatin clusters, obtained by multiple MD-runs, are shown for individual clusters of structures upon structural alignment and clustering on 5' (b) and 3' site (e) of G-quadruplex. Charge-charge contribution to interaction between (+)-ve amino groups on the ligand's side-chains, and (-)-ve phosphate groups of the G-quadruplex is evaluated by a 'charge-charge' score (*yellow*) for both 5' (a) and 3'site (d), and the mean RMSd values of the heavy atoms of the 5' (c) and 3' (f) site G-quartets of the individual clusters (≥ 3) are displayed in (*grey*).

6.5 CONCLUSIONS

A systematic and general approach to determining all plausible positions for binding a ligand to a G-quadruplex structure was developed. The flexibility of both the target and the ligand, as well as any conformational changes upon ligand binding were fully taken into consideration. Two structurally different compounds, rigid RHPS4 (with no rotatable bonds), and conformationally variable pyridostatin (with 13 rotatable bonds) were compared, both being biologically-active telomeric G-quadruplex binders, interacting with terminal G-quartet of a G-quadruplex through a stacking mode.

(I) Multiple starting positions for explicit solvent MD simulations were generated subsequent to placing the ligands in a grid-like manner and in multiple orientations, parallel to the ends of the quadruplex structure. This allowed for large areas of conformational space for both the ligand and its target, together with ligands transitions, to be explored. A total of 2436 starting structures (2204 pyridostatin/G-quadruplex complexes and 232 RHPS4/ G-quadruplex complexes) were employed, converging into a number of stable binding conformations rather than just one. Thus rather than a static model for binding, a fully dynamic model was utilized, providing an advantage over a simple docking approach.

(II) There was a good agreement (consistent energies) between the three methods used for binding free energy calculations (i.e semi-empirical PM6-DH2, and force field-based MM/PBSA and MM/GBSA). The averaged binding energies calculated for the rigid ligand RHPS4 were in better agreement than for the flexible pyridostatin molecule (with respect to the three methods employed). Notably higher correlations of the binding free energies predictions were found for the flexible pyridostatin molecule, when the individual methods were compared among each other. The highest correlation was found between the MOPAC and MM/PBSA predictions for pyridostatin on both 3’ and 5’ site of G-quadruplex, with a correlation coefficient of 0.88 and 0.92 respectively. Close correlation with a semiempirical method (i.e a higher level of theory) suggests a good level of accuracy of the force field-based method MM/PBSA, for free energy

calculations of a flexible ligand with strong electrostatic contribution to the overall interaction.

(III) No significant difference of calculated binding energies was found between the 3‘face and 5‘face of the G-quadruplex. In terms of the rigid RHPS4 molecule, the averaged binding energies of RHPS4 clusters bound with the 3‘face of G-quadruplex differed only by ~1 kcal/mol (more favourable) comparing to the 5‘face. However pyridostatin has shown an overall preference for the 5‘site stacking interaction with the G-quadruplex by ~ 4 kcal/mol difference when compared to the 3‘site.

(IV) The G-quadruplex grooves were not explored in this study, which could be seen as a limitation to the overall “picture”. Although based on experimental data, the end-stacking mode of interaction is relevant for these two ligands. This study has shown that the end stacking G-quadruplex ligands interact in multiple stable binding conformations on both the 3’- and 5’ face of the G-quadruplex, without a strong preference for a particular “energetically favourable” region.

CHAPTER 7:
Fragment-based design of G-quadruplex DNA ligands targeting c-MYC

7.1 BACKGROUND:

G-quadruplexes have been found to occur in tandem repeat regions of telomeres[242,244] in eukaryotic organisms and are abundant throughout the GC rich regions of the genome, particularly in gene promoters.[73,245] G-quadruplexes have been identified in the promoters in of several clinically-relevant oncogenes, notably *c-MYC*, *KRAS*, and *c-KIT*, amongst others.[76] Furthermore, it has been shown that G-quadruplexes that can form within the P1 promoter of c-MYC can regulate its transcription. Small molecules that bind and stabilize the c-MYC G-quadruplex have since been shown to suppress transcriptional activity in cancer cells.[261,262] Ongoing research in the groups of Neidle and Balasubramanian have shown the potential of small molecules to attenuate or promote the transcriptional activity of the c-KIT oncogene in human cells.[263] G-quadruplex ligands are not only valuable as tools for elucidating the mechanistic role of these secondary structural motifs, but also as possible therapies for a range of diseases that depend on the expression (and over-expression) of a particular gene.[264] The potential of G-quadruplex ligands in cancer therapy has been illustrated with CX-3543,[265] the first G-quadruplex interactive agent to enter human clinical trials, which disrupts nucleolin/ rDNA (i.e recombinant DNA) G-quadruplex in the nucleus.

An important challenge arising from the discovery of these biologically relevant G-quadruplexes is the necessity to identify drug-like small molecules that can discriminate not only between G-quadruplex and duplex DNA, but also between different G-quadruplexes.[99] This can be pursued through rational design, designing G-quadruplex binding small molecules based on what we know about the secondary structure of the G-quadruplex and using molecular modeling approaches. Knowledge of the crystal

structures of several complexes between human telomeric quadruplexes and substituted naphthalene diimide derivatives[79] has led to rational design of improved analogues with superior pharmacological properties (Micco *et al*, to be published). Another approach is to use smaller fragments to self-assemble around a labelled or tagged target, whereby bound fragments can be identified. Fragment-based methods have successfully been used to discover high affinity ligands for protein active sites, as well as target protein-protein interfaces. For example, Yang *et al*[266] used a thiol-based tethering strategy to screen 15.000 fragments to discover an inhibitor for BACE-1 (β-site amyloid precursor protein cleaving enzyme), a prominent target in Alzheimers disease. For nucleic acids, only a few examples of a fragment-based approach are available in the literature: most notably the recent efforts to target to target RNA riboswitches.[267] To date, however, targeting DNA by a fragment-based approach is still largely unexplored.

This work presents a fragment-based small molecule screen of the c-MYC G-quadruplex structure, which has been determined by the NMR methods (Nasiri *et al,* manuscript in preparation, 2012). Hit molecules, which are shown to bind c-MYC G-quadruplex structure in vitro and down-regulate cellular c-MYC in human HT1080 osteosarcoma cells, have been successfully discovered in Balasubramanian's research group (University of Cambridge), with a screening platform against c-MYC DNA, with an intercalator-displacement binding assay, where thiazole orange (TO) was used as an intercalator. TO is highly fluorescent when bound to target DNA and quenched after displacement (ΛEx= 501nm; Em= 539nm). In preliminary experiments, the dissociation constant (Kd) of TO to c-Myc DNA was determined (Kd = 3.5 μM ±0.69). Subsequently, the intercalator-displacement-assay (IDA) was applied in a high throughput fashion to screen an internal fragment library, which has been proven to be suitable for targeting RNA riboswiches against the G-quadruplex DNA target. The available fragment library of 1377 fragments were structurally and chemically diverse entries, compiled from various commercial sources, kindly provided by Professor Christopher Abell (University of Cambridge). Library compounds obey the 'rule of three', where; MW≤ 300 Da, clogP ≤ 3, no more than three hydrogen bond donors and acceptors. All fragments were ≥ 95% pure and have ≥ 1mM aqueous solubility. 15 hits were defined

through the statistically-robust assay, and the identified hit fragments were then subdivided in groups based on their chemical structure (Figure 7.1).

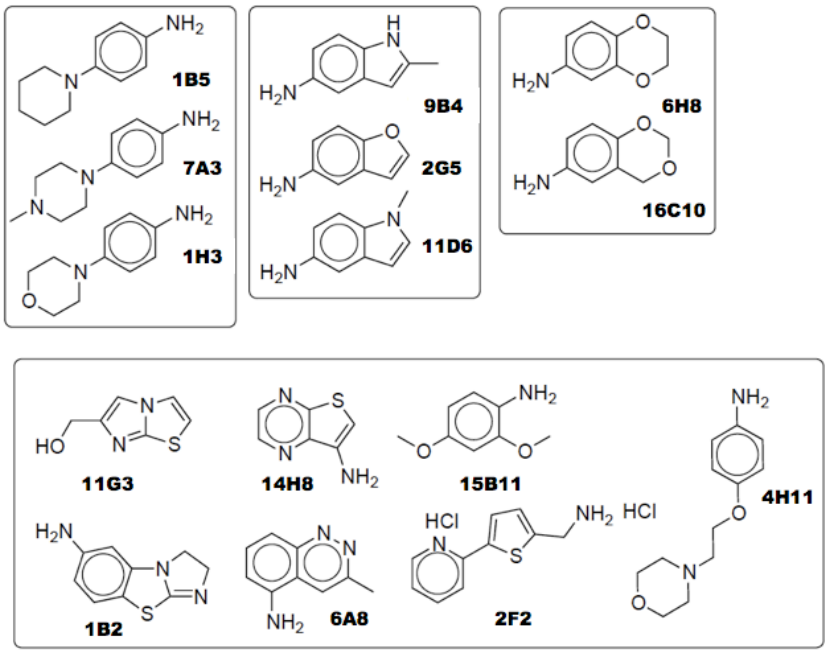


Figure 7.1: Top 15 fragment hits selected from the TO-intercalator displacement binding assay HTP screening against c-MYC DNA.
These 15 fragments were subsequently subjected to an *in silico* study using the NMR structure of biologically-relevant G-quadruplex element in human c-MYC promoter region (PDB id 1XAV).

7.2 AIMS:

From the 15 small-molecule (fragment) hits identified by screening an in-house library of fragments by means of an IDA assay (with TO employed as an intercalator), the best ten hits were further tested in a dose response titration experiment, and the best five fragments were then subjected to Biacore experiments to determine binding affinities (used as an independent secondary counter assay). All the experimental work outlined in this chapter was carried out by Balasubramanian's research group (University of Cambridge), as described by Nasiri *et al* (manuscript in preparation, 2012).

In parallel with this experimental work, an *in silico* study was performed with the aim to assess whether this could provide a reliable, more rapid and economic approach to finding hit fragments. For the *in silico* work, the NMR structure of the biologically relevant G-quadruplex element in the human c-MYC promoter[268] was used (PDB id 1XAV). The G-quadruplex adopts an intramolecular parallel-stranded quadruplex conformation with three guanine tetrads and three stable side loops, including two single-nucleotide side loops and one double nucleotide side loop, connecting the guanine strands.

7.3 METHODS:

The NMR structure of the biologically relevant G-quadruplex element in the human c-MYC promoter[268] (PDB id 1XAV) was retrieved from the PDB Database. A full length 22-mer d([TGA GGG TGG GTA GGG TGG GTA A]) and its 5' truncated version (the flanking dT5' removed; i.e 21-mer), were used for a molecular docking study with 15 best-ranked fragment hits identified by HTP screening (Figure 7.1, page 166). Subsequently, explicit solvent MD simulations were carried out for the 15 21-mer/fragment complexes, followed by MM/PB(GB)SA[131,132,229] calculations.

7.3.1 System setup and molecular docking

The 3D structures of the fragments were built by means of the ChemBio Office suite, and their conformations were optimized by a short cycle (500 steps) of MM2[269] energy minimization procedure (as MM2 is commonly used for small organic molecule calculations). With the exception of fragment 2F2 (net charge +2) their overall net charges were kept neutral. Suggested conformations of those fragments, with functional groups attached to their substituted cyclohexane ring, such as 4H11 and 7A3, were inspected by means of the Discovery Studio Visualizer program (http://accelrys.com). The equatorial position for their functional groups was verified, as it is sterically more plausible than the axial position. The 15 fragments were then docked with the energy-minimized G4-cMYC 22-mer (1) and the 5' truncated 21-mer (2) using the DOCK v 6.4[224] program. The entire surface of the G-quadruplexes was defined as a “binding site” (all the spheres generated by the *sphgen* program of DOCK 6, representing the binding site were employed (i.e cluster 0) to allow all possible binding poses of the small molecule ligands to be examined. As previously described in the DOCK 6 docking protocol (CHAPTER 5, section 5.3.3), the *anchor-and-growth* strategy for incremental ligand construction, allowing for the ligand's flexibility, was employed. Grid-based (primary) and the Hawkins GBSA[230,231] (secondary) scoring functions were subsequently used to rank the three best ligand's orientations, with the highest-ranked binding pose of each fragment being further examined.

7.3.2 Molecular dynamics simulation and MM/PB(GB)SA calculations

5-ns molecular dynamics simulations were performed for the 21-mer alone (as a reference), and for the 15 21-mer/fragment complexes with the best binding poses of the ligands suggested by molecular docking. All full-atom simulations were performed with the GROMACS v 4.5.3 program, employing the *parmbsc0*[161] force field previously ported into GROMACS. The topologies and other parameters for the small-molecule fragments were obtained via the ACPYPE[169] tool, employing the ANTECHAMBER[170] module of the AMBER11 program with the GAFF[226] force field. All MD protocols were kept identical for consistency of the results. Explicit solvent simulations were performed at T=300K with a time constant for coupling of 0.1 ps under the control of a velocity rescaling thermostat,[199] and isotropic constant-pressure boundary conditions controlled by the Parinello-Rahman[200,201] algorithm of pressure coupling. Long-range electrostatics were calculated using the PME[122] algorithm with grid spacing of 1.17 Å, and the LINCS[126] algorithm was employed to constrain all bonds. Non-bonded van der Waals interactions were treated in terms of Lennard-Jones 12-6 potential with a 10.0 Å cutoff. The solute was soaked in a triclinic box of TIP3P[196] water with a minimal clearance of 20.0 Å between periodic images for the starting configurations. Additionally, positively-charged K+ counter-ions were included in the systems to neutralize the negative net-charge on the DNA backbone. In each of the MD runs, there were two temperature-coupling groups; DNA with the structural K+ ions (and ligand, when present), and water with counter-ions. Subsequently, the systems were subjected to 10,000 steps of potential energy minimization, followed by 300 ps of molecular dynamics at 200K while keeping the solutes constrained, and further 100 ps of MD during which the systems were slowly heated to 300K and further equilibrated prior to unconstrained 5-ns production-level MD trajectory calculations. The time-step applied was 2.0 fs with coordinates saved every 5.0 ps. The initial 500 ps were then rejected for the subsequent MM/PB(GB)SA calculations, that were carried out over 450 frames representing the last 4.5 ns of the 5-ns production runs of the 21-mer/fragment complexes. A corresponding protocol previously described in CHAPTER 6, methodology section 6.3.3. was applied here.

7.4 RESULTS AND DISCUSSION:

The bound fragment positions found by the docking for the truncated c-MYC 21-mer (T5' removed) are shown in Figure 7.2, together with schematic drawing of the native c-MYC 22-mer (PDB id 1XAV). A corresponding numbering of quadruplex bases was used (Figure 7.2 a). The truncated 21-mer was of greater relevance comparing to the 22-mer, as the former was also employed in the experimental part of this study. Thus only the results concerning the 21-mer will be discusses further. All fragments, with the exception of 7A3, were found to be docked within the T14-A15 loop region. Fragment 7A3 docked within the groove formed between the third G4-tetrad (G9-G13-G18-G22) and the 3'-terminal residues stacking over the top of the 3rd G4-tetrad (Figure 7.2 b,c). This finding is in accord with the experimental cell-based studies, where different behavior of fragment 7A3 was observed compared to the rest of the 15 fragments.

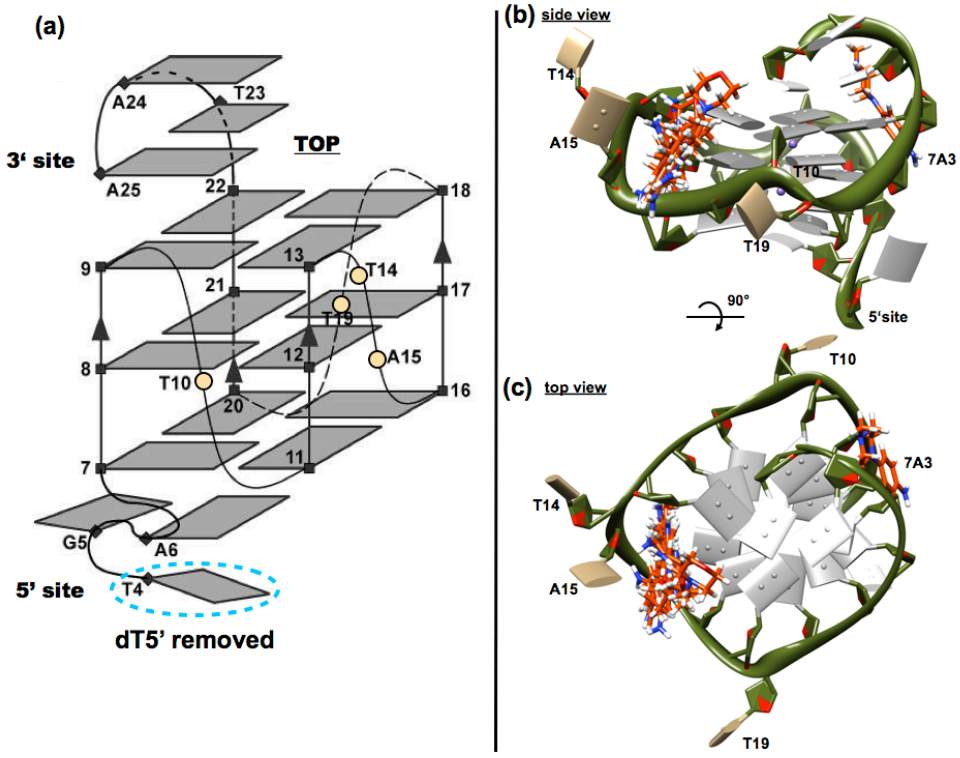


Figure 7.2: Schematic views of the c-MYC 22-mer G-quadruplex, and binding poses of the 15 fragments with the truncated 21-mer, found by the DOCK6 procedure.
(a) the cMYC 22-mer schematic drawing; strand directionality is indicated by the arrows, and the loops-forming nucleotides are highlighted by *beige* circles (Adapted from Ambrus *et al*., 2004). The 21-mer was formed by deletion of the 5'-thymine as shown. (b,c) Two views of the 21-mer with binding poses of the fragments found by DOCK6, in ribbon representation (*green*) with guanines in *grey*, and loop-forming nucleotides highlighted in *beige*. The fragments in stick representation are colored *orange*.

In terms of the structural stability observed throughout the 16 (15 complexes and the reference native 21-mer structure) 5-ns MD runs, the G-quadruplex structures remained entirely intact (with the structural K+ ions present in the channel) for all complex-bound (21-mer/fragment complex) and the reference (21-mer alone) model. RMSD values for the all-atom 21-mer/fragment complexes, and the complex-bound 21-mers respectively, as a function of simulation time were used as a measure of stabilization of the 15 complexes. Correspondingly, RMSD values for the all atom reference 21-mer (the uncomplexed 21-mer) were calculated. The simulation for this structure stabilized at ~2.4 Å, and ~1.4 Å respectively for the time-averaged RMSD plot (Figure 7.3 a). However, whereas majority of the ligands remained at their initial binding site (the T14-A15 loop), or at its vicinity, throughout the MD run, fragments 2G5, 11D6 and and eventually 9B4 left the binding site completely, and 'escaped'. Also fragment 2F2 was found to be moving away from binding site toward the end of the 5-ns simulation time, while fragment 1B2 only flipped-out of the binding site in the last 1 ns of the 5-ns MD run. Two fragments, 1B5 and 6A8 were found to relocate from their initial binding site on top of the 3rd G4-tetrad formed by G9-G13-G18-G22, with the latter stacking with G18 (Figure 7.3 b, c, Table 7.1).

The flexible regions of the 21-mer G-qudruplexes (16 in total, one reference 21-mer and 15 complex-bound) were also analyzed by examining their structural fluctuations in terms of the RMSF as a function of residue number (Figure 7.3 d). Overall, the G-tetrads have are very stable, with most of the overall structural flexibility (indicated by peaks) in the quadruplexes in large part arising from the loops. However, these fluctuations are relatively marginal, reflecting the nature of the stable, single- (T10 and T19) and double-nucleotide (T14-A15) loops. All 15 fragments showed a stabilizing effect on the T14-A14 loop region where they were initially docked, when compared to the reference 21-mer (Figure 7.3 d). The binding of fragments 6H8, 4H11 or 15B11 significantly reduced loop flexibility. In contrast, the 5'-flanking G5-A6 sequence show increased flexibility in almost all of the 21-mer/fragment complexes (except with fragment 14H8) compared to the native c-MYC 21-mer. The 3'-flanking T23-A24-A25 sequence is stacked over the top of the 3rd G4-tetrad in all 16 models, showing some

structural flexibility, suggesting that the 3'-end maintains a stable conformation over the course of the 5-ns simulation.

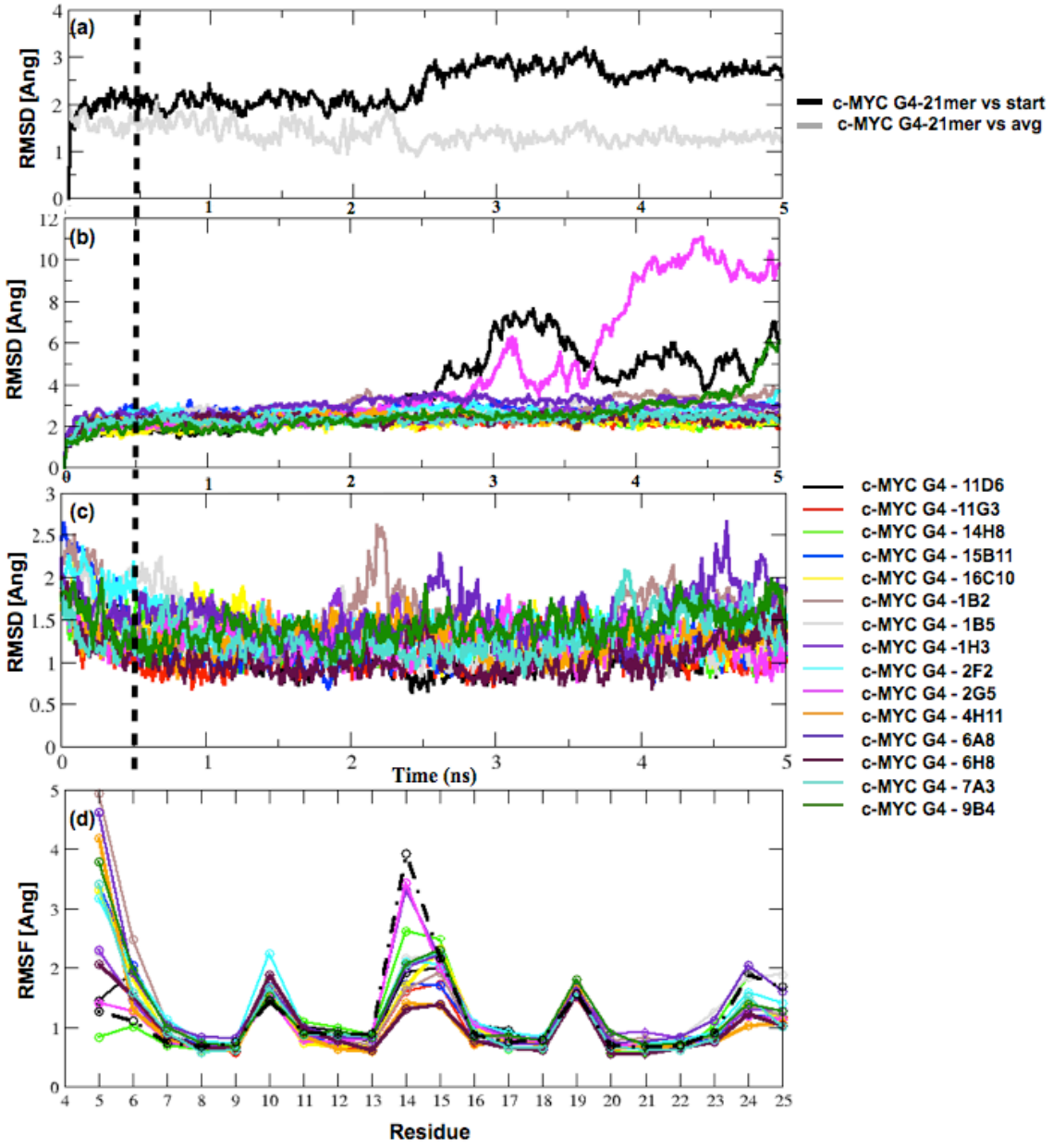


Figure 7.3: RMSD and RMSF plots showing the stability of the simulated systems during the MD simulations of the c-MYC 21-mer/fragment complexes.
(a) all-atom RMSD of the truncated 21-mer with respect to the initial (black) and the time-averaged (grey) structure; (b) all-atom RMSD plots of the G4/fragment complexes with respect to their starting structures, and (3) all-atom RMSD of the complex-bound 21-mer with respect to the time-averaged structures over the 5 ns MD run; Fragments 2G5, 11D6 and 9B4 'escape' the binding site; (d) RMSF per residue plots of the complex-bound 21-mer with respect to the initial structure. The truncated 21-mer (alone) is shown in black dotted line, to demonstrate the fragment's stabilizing effect.

Relative binding energies for each of the 21-mer/fragment complexes were calculated from the MD trajectories by means of the MM/PB(GB)SA program (AMBER11 package). MM/PB(GB)SA computes the binding energies employing the thermodynamic cycle that combines molecular mechanics (MM) energies with implicit solvent methods (Table 7.1). The initial 0.5 ns of each trajectory were rejected, and the last 4.5 ns represented by 450 frames were used for the calculations. Plausible hydrogen bonds formed between the individual fragments and the truncated c-MYC 21-mer were initially determined by means of the UCSF Chimera program upon docking the fragments with the G-quadruplex target (this is the 'static state', using the best ranked fragment pose upon docking). Hydrogen bonds formation throughout the short MD runs was also mapped with the VMD molecular visualization program (with hydrogen bond distance cutoff 3.2 Å). Only hydrogen bonds lasting at least 0.5 ns were considered (Table 7.1)

The overall score for each fragment was assessed, by ranking each according to its predicted binding energy (MM/PBSA and MM/GBSA) and stability from 1 (best) to 15 (worst), and combining their scores. Hydrogen bonds formed between the G4/fragment complexes were also considered within the overall score (Table 7.2). Overall, fragments 6H8 and 16C10 performed the best with consistent predicted binding energy and complex stability throughout the MD trajectories. These gave improved G-quadruplex stabilization, hydrogen bond formation and strongly favourable binding energies (with PB and GB values in good agreement). Fragments 7A3, 4H11 and 15B11 also ranked towards the high end of the group. At the other side of the ranking scale, 2G5 and 11D6 did not appear to perform well since their intermolecular binding interactions were significantly less favourable than any other of the 15 fragment studies, and moreover, both completely left the binding site during the course of 5-ns MD runs.

Table 7.2: Fragments ranked according to their binding energies and stability plots, over the 15 5-ns MD simulations.

[kcal/mol]															
GB	15B11	6A8	6H8	14H8	11G3	16C10	1B2	1B5	4H11	9B4	7A3	2F2	1H3	2G5	11D6
PB	7A3	6H8	4H11	16C10	9B4	2F2	1H3	15B11	1B2	6A8	1B5	11G3	14H8	11D6	2G5
ORDER	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
RMSD (complex)	16C10	11G3	14H8	6H8	7A3	4H11	1H3	9B4	2F2	1B2	1B5	15B11	6A8	11D6	2G5
RMSD (DNA)	16C10	11G3	2G5	14H8	6H8	11D6	9B4	1B5	4H11	7A3	1H3	2F2	1B2	6A8	15B11
RANK	**6H8**	**16C10**	**7A3**	**4H11**	**15B11**	**11G3**	**14H8**	**6A8**	**9B4**	**1B2**	**1H3**	**1B5**	**2F2**	**11D6**	**2G5**

Table 7.1: Overview of the results of the cMYC 21mer *in silico* study with 15 small-molecule fragments.
Hydrogen bonds (HB) suggested upon docking, as well as through the course of the MD runs are listed. The RMSd values for the entire G-quadruplex/fragment complexes, and the complex-bound 21mer only are summarized. Relative binding intermolecular energies were calculated via both GB and PB implicit solvent methods. (The fragments are listed according to their structural similarity, and color-coding corresponds to Figure 7.3)

FRAG-MENT	**HB (docked)**	**HB (trajectory)**	**fragment's location**	**RMSD [Å]**		**GB/SA score**	**MM/PB(GB)SA [kcal/mol]**		**score**
	(UCSF chimera)	(VMD, cutoff 3.2 Å)	(MD 5-ns)	**G4/fragment**	**G4 only**	[kcal/mol]	GB	PB	**Ene; rms**
1B5	G13_O4'	---	T14-A15 loop → 3'site	2.69	2.32	-19.24	-10.58	-9.71	19; 19
7A3	G8_OP	G8; G22	groove	2.33	2.35	-20.25	-9.28	-13.87	12; 15
1H3	G16_OP	G17	T14-A15 loop	2.44	2.42	-20.79	-8.46	-10.89	20; 18
9B4	G13_O4'	G12; G13	escapes	2.51	2.27	-20.41	-9.51	-11.37	15; 15
2G5	G16_OP	---	escapes	4.55	2.18	-19.67	-2.75	-1.47	29; 18
11D6	G16_OP	---	escapes	3.73	2.26	-20.84	-2.64	-1.98	29; 20
6H8	G16_OP	G12	T14-A15 loop	2.32	2.26	-20.35	-11.85	-12.70	5; 9
16C10	G13_N2; G18_OP	G12; G13	T14-A15 loop	2.10	2.07	-18.40	-11.21	-11.76	10; 2
14H8	G12_N2; G13_O3'; A15_N3	G13; G17	T14-A15 loop	2.25	2.19	-18.64	-11.74	-9.06	17; 7
11G3	G12_N2; G16_OP	G12; G16	T14-A15 loop	2.19	2.15	-18.89	-11.46	-9.44	17; 4
15B11	G12_N2; G13_O3'; T14_OP	G12; G13	T14-A15 loop	2.71	2.74	-20.51	-12.41	-10.56	9; 27
6A8	G13_N2; A15_O3'	G18	T14-A15 loop → 3'site	2.97	2.70	-20.97	-12.20	-10.08	12; 27
1B2	G13_N2; G17_OP	G13	T14-A15 loop *	2.66	2.60	-21.08	-10.83	-10.37	16; 23
2F2	G16_OP	G13; T14	escapes	2.63	2.51	-22.64	-8.69	-11.00	18; 12
4H11	G16_OP; G12_N2	G13; A15; G17	T14-A15 loop	2.38	2.33	-24.53	-9.97	-12.53	12; 15

Binding poses of the best ranked fragments 6H8, 16C10 and 7A3 are shown in Figure 7.4 in detail, both upon docking, and at the end of their 5-ns MD runs (last frame of the 5-ns trajectory was used). This finding is in accord with the experimental work, as fragments 6H8 and 16C10 were identified among the four best inhibitors (16C10, 11D6, 6H8 and 14H5, Figure S7.1 in Supplementary section), reducing c-MYC in all repeat experiments, with fragment 6H8 being the top hit identified both experimentally and *in silico*. It is worth mentioning, that the two best-ranked fragments are structurally very similar (Figure 7.1), with oxygen atoms in their di-substituted cyclohexane ring being either in para (6H8) or meta (16C10) positions; these were then found to be involved in specific hydrogen-bond formation with N2 of G12. Similarly, 2G5 and 11D6 at the very end of the ranking spectra are structurally alike, and together with fragment 9B4 they all moved away from their binding site through the course of the MD run. Structurally-similar fragments 1H3 and 1B5 also ranked towards the lower end of the group, however fragment 7A3, which is also structurally similar to 1H3 and 1B5, performed well in terms of stability and binding energy. The difference may lay in an extra methyl group at the para position of the disubstituted cyclohexane ring, which may contribute towards its preferable binding properties at the G-quadruplex groove.

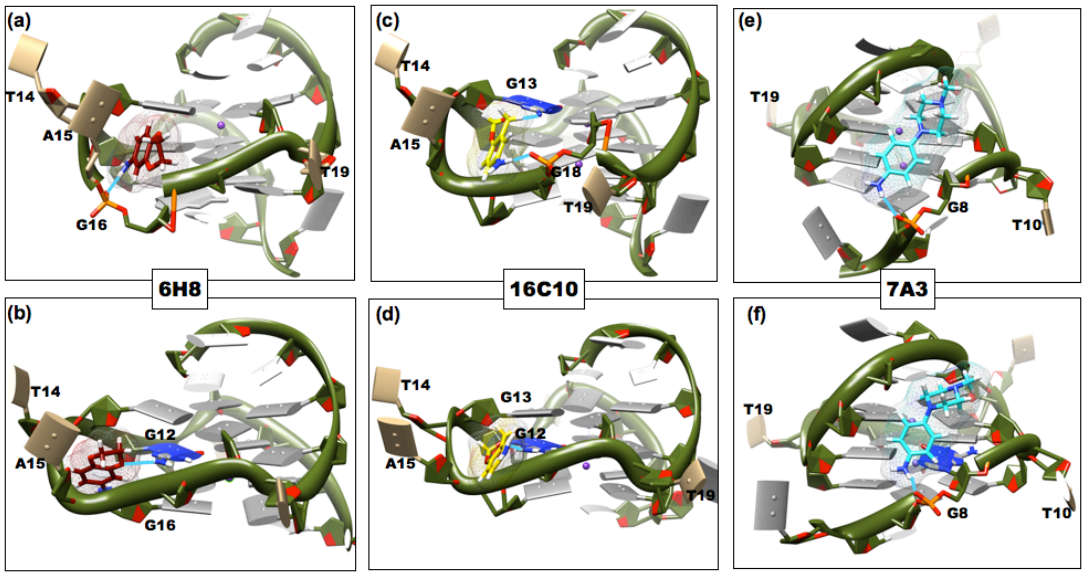


Figure 7.4 : Best predicted fragments 6H8, 16C10 and 7A3 shown bound to the G4-cMYC 21mer. The predicted binding poses of fragments upon docking are shown in panels (a, c, e), and at the end of the 5-ns MD runs in panels (b, d, e); Graphical representation of the G-quadruplex is corresponding to Figure 7.2, with the fragment shown in stick representation, and colored *brown* (6H8), *yellow* (16C10) and *turquoise* (7A3) and by heteroatom. (Fragments's colors also correspond to figure 7.3). Hydrogen bonds formed between the fragments and 21-mer bases are shown in *cyan*.

7.5 CONCLUSIONS:

An *in silico* study focused on fragment-based design of G-quadruplex DNA ligands targeting c-MYC was performed in parallel with an experimental approach, in order to assess the feasibility of providing a more rapid, reliable and economic approach to finding hit fragments. The 15 fragments that were employed have been previously screened and determined by NMR methods as being able to bind c-MYC G-quadruplex structure *in vitro,* and they were shown to down-regulate cellular c-MYC in human HT1080 osteosarcoma cells.

(I) A combination of several computational tools were applied to identification of a promising hit, namely (in a sequential manner) molecular docking of the fragments with a truncated 21-mer c-MYC G-quadruplex, explicit solvent MD simulations of the individual proposed complexes, and their free binding energy calculations, provided a valuable insight into the dynamic behavior and stability of these complexes.

(II) The docking-proposed binding poses of the fragments with the 21-mer c-MYC were in excellent agreement with the experimental results, where only fragment 7A3 was observed bound elsewhere than the rest of the fragments.

(III) The combined *in silico* approaches then led to a successful prediction of two best-ranked fragments, 6H8 and 16C10, that were also among the best experimentally determined compounds, with an ability to significantly reduce c-MYC in all repeat experiments. Fragment 6H8 was ranked as the best inhibitor in both experimental, and *in silico* study.

(IV) Employing multiple computational tools simultaneously in hit identification was shown to provide an advantage to using a single technique, such as rigid receptor docking in this case. If only docking was employed here, fragments 4H11, 2F2 or 1B2 would have been suggested as promising inhibitors (based on the GB/SA scores), neither of them being on the list of most-successful inhibitors selected experimentally.

Supplementary information for CHAPTER 7:

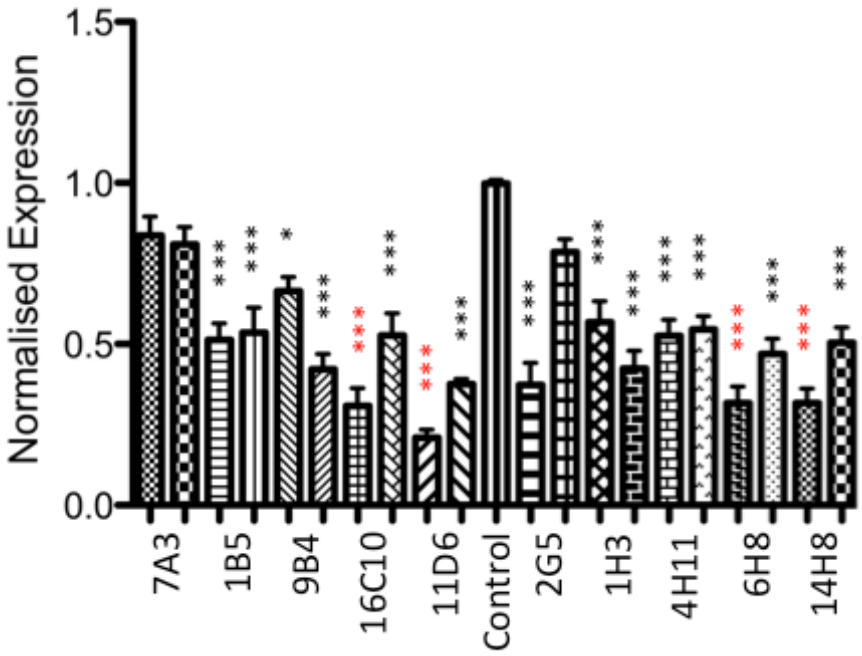


Figure S7.1: c-MYC experimental data.[11]
The target c-MYC protein translation was investigated in presence of hit fragments by quantification of c-MYC expression using a c-MYC specific antibody. Combined data from five plates, at least 18 data points per well. Statistical significance per ANOVA (Kruskall-Wallis test; $P<0.0001$). Four treatments significantly reduced c-MYC in all repeat experiments; these were 16C10, 11D6, 6H8 and 14H5, of which 6H8 was the best inhibitor. These fragments were used to carry out further binary treatments, whereupon all binary treatments (125 mM of each component) induced a significant reduction (ANOVA). The best binary mixture was 6H8*11D6.

[11] The experimental data was kindly provided by Dr. Hamid Nasiri and Professor Shankar Balasubramanian

‘FINALE’

CONCLUDING REMARKS

Molecular modeling and computational chemistry methods have become profoundly integrated into the drug discovery process over the last quarter century. Indeed, *in silico* approaches can provide various levels of insight into the drug-target (ligand-target) binding and recognition, which require application of different computational methods along the drug discovery pipeline. In the earliest stages of the drug discovery process, homology modeling and molecular dynamics simulations (or Monte Carlo sampling) can contribute to target validation, and target's conformational sampling respectively. Virtual screening (structure-based or ligand-based) for potential small molecule ligands can be then applied at the hit identification stage, together with 3D-pharmacophore modeling. Free energy calculations of ligand-receptor binding are then another natural application of *in silico* approaches in drug discovery. Further explicit solvent molecular dynamics simulations of the receptor-ligand complexes may be applied in the proceeding stages of the pipeline to complement and rationalize experimental observations in the absence of structural data. Molecular modeling (and computational chemistry) methods may both challenge and complement current experimental techniques by providing spatial and temporal resolutions to allow for in-depth analysis and further understanding of the mechanisms underlying the biological and chemical systems.[270]

Molecular dynamics simulations are being increasingly mated with numerous experiments, as simulations may trace system behavior over large "spatiotemporal domain-length scales" (*i.e* atomic precision, femtosecond resolution, and timescales up to microseconds).[270] Together with the computational power, that has tremendously increased in the last two decades or so (with parallel capability of current computer architectures), the amount of produced data (i.e trajectories) is proportionally increasing. However, effective post-processing analysis tools and methods that can efficiently deal with the volume of the data are yet needed. There has also been a continuous improvement in the simulation algorithms and parameters, but their current limitations (i.e the simulation time scale that is not biochemically relevant to biomolecular dynamics, and the approximate nature of the biomolecular force fields) should be always kept in mind to avoid mis-interpretation of the simulation results. Whereas more approximate, empirical, methods are adequate for large biomolecular systems simulations, more delicate questions that require the quantum electronic effects to be accounted for (i.e bonds

forming/breaking, polarization effects or charge transfer) need to be answered at higher level of theory, by means of quantum mechanics-based approaches.

MD simulations served as the imaginary "spine" of this thesis, as they were applied, and purposefully tailored to all studied systems that are described here. In first part of the thesis, explicit solvent full atom molecular dynamics simulations of the STAT3βtc homodimer:DNA complex were performed, in both its phosphorylated and unphosphorylated form, as well as in complex-bound and latent monomeric form. Valuable insight into the aspects of the recognition at both protein-DNA-solvent, and protein-protein level was reached, together with structural explanation of experimentally determined point mutations. The modeling results provided a theoretical platform for X-ray studies of the unphosphorylated STAT3-DNA complex, supporting the recently emerged evidence STAT3 non-canonical signaling (Nkansah *et al*, manuscript submitted for publication, 2012). A number of point mutations of the residues involved in protein-DNA hydrogen bond formation have been also examined experimentally, by means of X-ray crystallography (Parkinson *et al*, work in progress), aiming to explore the 3D structural consequences of naturally occurring disease-related mutations. Furthermore, single-site mutations of STAT3β proteins will be generated to probe the conformational surface of the SH2 domain, in the presence or absence of selected ligands. Multiple receptor conformations obtained from the MD trajectory were directly implemented into *in silico* approaches to the discovery of novel STAT3-STAT3 inhibitors for chemotherapeutic intervention, employing a comparative multiple receptor conformation molecular docking study, providing an advantage over more traditional single receptor conformation studies, as the dynamic aspect of the receptor behavior was introduced here.

A truly dynamic (fully flexible) approach to receptor-ligand binding was taken in the second part of the thesis. A systematic and general approach to determining all plausible positions for binding a ligand to a G-quadruplex structure was developed. The flexibility of both the target and the ligand, as well as any conformational changes upon ligand binding were fully taken into consideration. Free binding energies of the vast number of ligand binding poses were then evaluated by means of semiempirical and

empirical methods, showing a very good correlation of these methods when the conformationally variable ligand was considered. In the last chapter, combined computational techniques were applied to fragment-based approach towards G-quadruplex stabilizing ligands, that was performed in parallel to ongoing experimental work. The combined *in silico* approaches led to a successful prediction of two best-ranked fragments, that were also among the best experimentally determined compounds, with an ability to significantly reduce c-MYC.

[1] Watson, J.D. and Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **1953**, 171, 737-738.

[2] Watson, J.D. and Crick, F.H. Genetical Implications of the Structure of Deoxyribonucleic Acid; *Nature* **1953**, 171, 964–967.

[3] Franklin, R.E. and Gosling, R.G. Evidence for 2-Chain Helix in Crystalline Structure of Sodium Deoxyribonucleate; *Nature* **1953**, 172, 156–157.184

[4] Huppert, J. Structure, Location and Interactions of G-quadruplexes; *FEBS J.* **2010**, 277, 3452–3458.

[5] Kolář, M., Kubař, T., Hobza, P. On the Role of London Dispersion Forces in Biomolecular Structure Determination; *J Phys Chem B* **2011**, 115, 8038–8046.

[6] Černý, J. and Hobza, P. Non-covalent Interactions in Biomacromolecules; *Phys Chem Chem Phys* **2007**, 9, 5291–5303.

[7] Hobza, P. and Šponer, J. Structure, Energetics, and Dynamics of the Nucleic Acid Base Pairs: Nonempirical Ab Initio Calculations; *Chem Rev* **1999**, 99, 3247–3276.

[8] Dickerson, R.E., Drew, H.R., Conner, B.N., Wing, R.M., Fratini, A.V., Kopka, M.L. The Anatomy of A-, B-, and Z-DNA; *Science* **1982**, 216, 475–485.

[9] Schlick, T. Nucleic Acids Structure Minitutorial in *Molecular Modeling and Simulation: An Interdisciplinary Guide*; ed. Schlick, T.; Springer New York **2010**, 26, 129–162.

[10] Neidle, S. DNA Minor-groove Recognition by Small Molecules (up to 2000); *Nat Prod Rep* **2001**, 18, 291–309.

[11] Chaires, J.B. Drug-DNA Interactions; *Curr Opin Struct Biol* **1998**, 8, 314–320.

[12] Rahman, K.M., James, C.H., Thurston, D.E. Effect of Base Sequence on the DNA Cross-linking Properties of Pyrrolobenzodiazepine (PBD) Dimers; *Nucleic Acids Res* **2011**, 5800–5812.

[13] Wemmer, D.E., Dervan, P.D. Targeting the Minor Groove of DNA; *Curr Opin Struct Biol* **1997**, 7, 355–361.

[14] Wang, Y. and Patel, D.J. Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex; *Structure* **1993**, 1, 263–282.

[15] Parkinson, G.N., Lee M.P., Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA; *Nature* **2002**, 417, 876-880.

[16] Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K., Neidle, S. Quadruplex DNA: Sequence, Topology and Structure; *Nucleic Acids Res* **2006**, 34, 5402–5415.

[17] Westhof, E. Water: An Integral Part of Nucleic Acid Structure; *Ann Rev Biophys Biophys Chem* **1988**, 17, 125–144.

[18] Schneider, B. and Berman, H. Basics of Nucleic Acid Structure in *Computational Studies of RNA and DNA*; ed. Šponer, J. and Lankaš, F.; Springer Netherlands **2006**, 1-44.

[19] Schildbach, J.F., Karzai, A.W., Raumann, B.E., Sauer R.T. Origins of DNA-binding specificity: Role of protein contacts with the DNA backbone; *Proc Natl Acad Sci* **1999**, 96, 811-817

[20] Sarai, A., Kono, H. Protein-DNA recognition patterns and predictions; *Annu Rev Biophys Biomol Struct* **2005**, 34, 379–398.

[21] Zhou, P., Tian, F., Ren, Y., Shang, Z. Systematic Classification and Analysis of Themes in Protein-DNA Recognition; *J Chem Inf Model* **2010**, 50, 1476-1488.

[22] Fletcher, S., Turkson, J., Gunning, P.T. Molecular Approaches Towards the Inhibition of the Signal Transducer and Activator of Transcription 3 (Stat3) Protein; *ChemMedChem* **2008**, 3, 1159–1168.

[23] Yap, J.L., Worlikar, S., MacKerell A.D.Jr, Shapiro, P., Fletcher, S. Small molecule Inhibitors of the ERK Signaling Pathway: Towards Novel Anticancer Therapeutics; *ChemMedChem* **2011**, 6, 38–48.

[24] Peukert, S. and Miller-Moslin, K. Small-molecule Inhibitors of the Hedgehog Signaling Pathway as Cancer Therapeutics; *ChemMedChem* **2010**, 5, 500–512.

[25] Yap, J.L., Chauhan, J., Kwan-Young, J., Chen, L., Prochownik, E.V., Fletcher, S. Small-molecule Inhibitors of Dimeric Transcription Factors: Antagonism of Protein–protein and protein–DNA Interactions; *MedChemComm* **2012**, 3, 541–551.

[26] Redell, M.S. and Tweardy, J.D. Targeting Transcription Factors in Cancer: Challenges and Evolving Strategies; *Drug Dis Today Technol* **2006**, 3, 261–267.

[27] Redmond, A.M. and Carroll, J.S. Defining and Targeting Transcription Factors in Cancer; *Genome Biol* **2009**, 10, 311.

[28] Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., Salzberg, A.C., Huang, E.S. Structure-based Maximal Affinity Model Predicts Small-molecule Druggability; *Nat Biotechnol* **2007**, 25, 71–75.

[29] Sioud, M. and Leirdal, M. Druggable Signaling Proteins; *Methods Mol Biol* **2007**, 361, 1–24.

[30] Darnell, J.E.Jr. STATs and Gene Regulation; *Science* **1997***,* 277, 1630–1635.

[31] Lim, C.P. and Cao, X. Structure, Function, and Regulation of STAT Proteins; *Mol BioSyst* **2006**, 2, 536–550.

[32] Turkson, J. STAT Proteins as Novel Targets for Cancer Drug Discovery; *Expert Opin Ther Targets* **2004**, 8, 409–422.

[33] Hirano, T., Ishihara, K., Hibi, M. Roles of STAT3 in Mediating the Cell Growth, Differentiation and Survival Signals Relayed Through the IL-6 Family of Cytokine Receptors; *Oncogene* **2000**, 19, 2548–2556.

[34] Zhang, X. and Darnell, J.E. Functional Importance of Stat3 Tetramerization in Activation of the Alpha 2-macroglobulin Gene; *J Biol Chem* **2001**, 276, 33576–33581.

[35] Zhang, T., Kee, W.H., Seow, K.T., Fung, W., Cao, X. The Coiled-Coil Domain of Stat3 Is Essential for Its SH2 Domain-Mediated Receptor Binding and Subsequent Activation Induced by Epidermal Growth Factor and Interleukin-6; *Mol Cell Biol* **2000**, 20, 7132–7139.

[36] Ma, J., Zhang, T., Novotny-Diermayr, V., Tan, A. L., Cao, X. A Novel Sequence in the Coiled-coil Domain of Stat3 Essential for Its Nuclear Translocation; *J Biol Chem* **2003**, 278, 29252–29260.

[37] Horvath, C.M., Wen, Z., Darnell, J.E. A STAT Protein Domain That Determines DNA Sequence Recognition Suggests a Novel DNA-binding Domain; *Genes Devel* **1995**, 9, 984–994.

[38] Yu, Z. and Kone, B.C. The STAT3 DNA-binding Domain Mediates Interaction with NF-kappaB P65 and Inducible Nitric Oxide Synthase Transrepression in Mesangial Cells; *J Am Soc Nephrol* **2004**, 15, 585–591.

[39] Paulson, M., Pisharody, S., Pan, L., Guadagno, S., Mui, A.L., Levy, D.E. Stat Protein Transactivation Domains Recruit p300/CBP Through Widely Divergent Sequences; *J Biol Chem* **1999**, 274, 25343–25349.

[40] Levy, D.E. and Lee, C. What does Stat3 do?; *J Clin Invest* **2002**, 109, 1143–1148.

[41] Yuan, Z., Guan, Y., Chatterjee, D., Chin, Y.E. Stat3 Dimerization Regulated by Reversible Acetylation of a Single Lysine Residue; *Science* **2005**, 307, 269–273.

[42] Schaefer, T.S., Sanders, L.K., Park, O.K., Nathans, D. Functional Differences Between Stat3alpha and Stat3beta; *Mol Cell Biology* **1997**, 17, 5307–5316.

[43] Schaefer, T.S., Sanders, L.K., Nathans, D. Cooperative Transcriptional Activity of Jun and Stat3 Beta, a Short Form of Stat3; *Proc Nat Acad Sci* **1995**, 92, 9097–9101.

[44] Caldenhoven, E.T., van Dijk, B., Solari, R., Armstrong, J., Raaijmakers, J.A., Lammers, J.W., Koenderman, L., de Groot, R.P. STAT3beta, a Splice Variant of Transcription Factor STAT3, Is a Dominant Negative Regulator of Transcription; *J Biol Chem* **1996**, 271, 13221–13227.

[45] Maritano, D., Sugrue, M.L., Tininini, S., Dewilde, S., Strobl, B., Fu, X., Murray-Tait, V., Chiarle, R., Poli, V. The STAT3 Isoforms Alpha and Beta Have Unique and Specific Functions; *Nat Immunol* **2004**, 5, 401–409.

[46] Yu, H., Kortylewski, M., Pardoll, D. Crosstalk Between Cancer and Immune Cells: Role of STAT3 in the Tumour Microenvironment; *Nat Rev Immunol* **2007**, 7, 41–51.

[47] Yu, H., Pardoll, D., Jove, R. STATs in Cancer Inflammation and Immunity: a Leading Role for STAT3; *Nat Rev Cancer* **2009**, 9, 798–809.

[48] Lee, H., Pal, S.K., Reckamp, K., Figlin, R.A., Yu, H. STAT3: A Target to Enhance Antitumor Immune Response; *Curr Top Microbiol Immunol* **2011**, 344, 41–59.

[49] Page, B.D., Ball, D.P., Gunning, P.T. Signal Transducer and Activator of Transcription 3 Inhibitors: a Patent Review; *Expert Opin Ther Patents* **2011**, 21, 65–83.

[50] Braunstein, J., Brutsaert, S., Olson, R., Schindler, C. STATs dimerize in the absence of phosphorylation; *J Biol Chem* **2003**, 278, 34133−34140.

[51] Yue, P. and Turkson, J. Targeting STAT3 in cancer: how successful are we?; *Expert Opin Invest Drugs* **2009**, 18, 45−56.

[52] Jing, N. and Tweardy, D.J. Targeting STAT3 in cancer therapy; *Anticancer Drugs* **2005**, 16, 601−607.

[53] Turkson, J., Ryan, D., Kim, J.S., Zhang, Y., Chen, Z., Haura, E., Laudano, A., Sebti, S., Hamilton, A.D., Jove, R. Phosphotyrosyl Peptides Block Stat3-mediated DNA Binding Activity, Gene Regulation, and Cell Transformation; *J Biol Chem* **2001**, 276, 45443–45455.

[54] Ren, Z., Cabell, L.A., Schaefer, T.S., McMurray, J. Identification of a High-Affinity Phosphopeptide Inhibitor of Stat3; *Bioorg Med Chem Lett* **2003**, 13, 633–636.

[55] Chen, J., Bai, L., Bernard, D., Nikolovska-Coleska, Z., Gomez, C., Zhang, J., Yi, H., Wang, S. Structure-Based Design of Conformationally Constrained, Cell-Permeable STAT3 Inhibitors; *ACS Med Chem Lett* **2010**, 1, 85–89.

[56] Turkson, J., Kim, J.S., Zhang, S., Yuan, J., Huang, M., Glenn, M., Haura, E., Sebti, S., Hamilton, A.D., Jove, R. Novel Peptidomimetic Inhibitors of Signal Transducer and Activator of Transcription 3 Dimerization and Biological Activity; *Mol Cancer Ther* **2004**, 3, 261–269.

[57] Gunning, P.T., Katt, W.P., Glenn, M., Siddiquee, K., Kim, J. S., Jove, R., Sebti, S.M., Turkson, J., Hamilton, A.D. Isoform Selective Inhibition of STAT1 or STAT3 Homo-dimerization via Peptidomimetic Probes: Structural Recognition of STAT SH2 Domains; *Bioorg Med Chem Lett* **2007**, 17, 1875–1878.

[58] Gomez, C., Bai, L., Zhang, J., Nikolovska-Coleska, Z., Chen, J. Yi, H., Wang, S. Design, Synthesis, and Evaluation of Peptidomimetics Containing Freidinger Lactams as STAT3 Inhibitors; *Bioorg Med Chem Lett* **2009**, 19, 1733–1736.

[59] Jing, N., Li, Y., Xiong, W., Sha, W., Jing, L., Tweardy, D.J. G-quartet Oligonucleotides: a New Class of Signal Transducer and Activator of Transcription 3 Inhibitors That Suppresses Growth of Prostate and Breast Tumors Through Induction of Apoptosis; *Cancer Res* **2004**, 64, 6603–6609.

[60] Jing, N., Zhu, Q., Yuan, P., Li, Y., Mao, L., Tweardy, D.J. Targeting Signal Transducer and Activator of Transcription 3 with G-quartet Oligonucleotides: a Potential Novel Therapy for Head and Neck Cancer; *Molec Cancer Ther* **2006**, 5, 279–286.

[61] Zhu, Q. and Jing, N. Computational Study on Mechanism of G-quartet Oligonucleotide T40214 Selectively Targeting Stat3; *J Comput Aided Mol Des* **2007**, 21, 641–648.

[62] Turkson, J., Zhang, S., Mora, L.B., Burns, A., Sebti, S., Jove, R. A Novel Platinum Compound Inhibits Constitutive Stat3 Signaling and Induces Cell Cycle Arrest and Apoptosis of Malignant Cells; *J Biol Chem* **2005**, 280, 32979–32988.

[63] Song, H., Wang, R., Wang, S., Lin, J. A low-molecular-weight compound discovered through virtual database screening inhibits STAT3 function in breast cancer cells; *Proc Natl Acad Sci* **2005**, 102, 4700−4705.

[64] Siddiquee, K., Zhang, S., Guida, W.C., Blaskovich, M.A., Greedy, B., Lawrence, H.R., Yip, M.L., Jove, R., McLaughlin, M.M., Lawrence, N.J., Sebti, S.M., Turkson, J. Selective chemical probe inhibitor of STAT3, identified through structure-based virtual screening, induces antitumor activity; *Proc Natl Acad Sci* **2007**, 104, 7391−7396.

[65] Xu, X., Kasembeli, M.M., Jiang, X., Tweardy, B.J., Tweardy, D.J. Chemical probes that competitively and selectively inhibit STAT3 activation; *PLoS One* **2009**, 4, 1−12.

[66] Matsuno, K., Masuda, Y., Uehara, Y., Sato, H., Muroya, A., Takahashi, O., Yokotagawa, T., Furuya, T., Okawara, T., Otsuka, M., Ogo, N., Ashizawa, T., Oshita, C., Sachiko, T., Ishii, H., Akiyama, Y., Asai, A. Identification of a new series of STAT3 inhibitors by virtual screening; *ACS Med Chem Lett* **2010**, 1, 371−375.

[67] Siddiquee, K.A., Gunning, P.T., Glenn, M., Katt, W.P., Zhang, S., Schroeck, C., Sebti, S.M., Jove, R., Hamilton, A.D. An oxazole-based small-molecule STAT3 inhibitor modulates STAT3 stability and processing and induces antitumor cell effects; *ACS Chem Biol* **2007**, 2, 787−798.

[68] Bhasin, D., Cisek, K., Pandharkar, T., Regan, N., Li, C., Pandit, B., Lin, J., Li, P.K. Design, Synthesis, and Studies of Small Molecule STAT3 Inhibitors; *Bioorg Med Chem Lett* **2008**, 18, 391–395.

[69] Fuh, B., Sobo, M., Cen, L., Josiah, D., Hutzen, B., Cisek, K., Bhasin, D. *et al.* LLL-3 Inhibits STAT3 Activity, Suppresses Glioblastoma Cell Growth and Prolongs Survival in a Mouse Glioblastoma Model; *Br J Cancer* **2009**, 100, 106–112.

[70] Lin, L., Hutzen, B., Li, P.K., Ball, S., Zuo, M., DeAngelis, S., Foust, E. *et al*. A Novel Small Molecule, LLL12, Inhibits STAT3 Phosphorylation and Activities and Exhibits Potent Growth-Suppressive Activity in Human Cancer Cells; *Neoplasia* **2010***,* 12, 39–50.

[71] Zhang, X., Yue, P., Fletcher, S., Zhao, W., Gunning, P.T., Turkson, J. A Novel Small-molecule Disrupts Stat3 SH2 Domain-phosphotyrosine Interactions and Stat3-dependent Tumor Processes; *Biochem Pharmacol* **2010**, 79, 1398–1409.

[72] Zhang, X., Yue, P., Page, B.D., Li, T., Zhao, W., Namanja, A.T., Paladino, D., Zhao, J., Chend, Y., Gunning, P.T., Turkson, J. Orally Bioavailable Small-molecule Inhibitor of Transcription Factor Stat3 Regresses Human Breast and Lung Cancer Xenografts; *Proc Natl Acad Sci* **2012**, 109, 9623–9628.

[73] Todd, A.K., Johnston, M., Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA; *Nucleic Acids Res* **2005**, 33, 2901-2907.

[74] Neidle, S. Human telomeric G-quadruplex: The current status of telomeric G-quadruplexes as therapeutic targets in human cancer; *FEBS J* **2010**, 277, 1118-1125.

[75] Neidle, S. and Parkinson, G. Telomere maintenance as a target for anticancer drug discovery; *Nat Rev Drug Discov* **2002**, 1, 383-393.

[76] Balasubramanian, S., Hurley, L., Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?; *Nat Rev* **2011**, 10, 261-275.

[77] Campbell, N.H., Parkinson, G.N., Reszka, A.P., Neidle, S. Structural Basis of DNA Quadruplex Recognition by an Acridine Drug; *J Am Chem Soc* **2008**, 130, 6722-6724.

[78] Parkinson, G.N., Ghosh, R., Neidle, S. Structural basis for binding of porphyrin to human telomeres; *Biochemistry* **2007**, 46, 2390-2397.

[79] Collie, G.W., Promontorio, R., Hampel, S.M., Micco, M., Neidle, S., Parkinson, G.N. Structural Basis for Telomeric G-quadruplex Targeting by Naphthalene Diimide Ligands; *J Am Chem Soc* **2012**, 134, 2723–2731.

[80] Martínez, P. and Blasco, M.A. Role of shelterin in cancer and aging; *Aging Cell* **2010**, 9, 653-666.

[81] Martínez, P. and Blasco, M.A. Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins; *Nat Rev Cancer* **2011**, 11, 161-174.

[82] Neidle, S. and Read M.A. G-quadruplexes as therapeutic targets; *Biopolymers* **2001**, 56, 195-208.

[83] Fadrná, E., Špačková, N., Štefl, R., Koča, J., Cheatham, T.E.3rd, Šponer, J. Molecular Dynamics Simulations of Guanine Quadruplex Loops: Advances and Force Field Limitations; *Biophys J* **2004**, 87, 227–242.

[84] Šponer, J., Cang, X., and Cheatham, T.E. 3rd. Molecular Dynamics Simulations of G-DNA and Perspectives on the Simulation of Nucleic Acid Structures; *Methods* **2012**, 57, 25-39.

[85] Cang, X., Šponer, J., Cheatham, T.E. 3rd. Explaining the Varied Glycosidic Conformational, G-tract Length and Sequence Preferences for Anti-parallel G-quadruplexes; *Nucleic Acids Res* **2011**, 39,4499–4512.

[86] Cang, X., Šponer, J., Cheatham, T.E. 3rd. Insight into G-DNA Structural Polymorphism and Folding from Sequence and Loop Connectivity Through Free Energy Analysis; *J Am Chem Soc* **2011**, 133, 14270–14279.

[87] Štefl, R., Cheatham, T.E.3rd, Špačková, N., Fadrná, E., Berger, I., Koča, J., Šponer, J. Formation Pathways of a Guanine-quadruplex DNA Revealed by Molecular Dynamics and Thermodynamic Analysis of the Substates; *Biophys J* **2003**, 85, 1787–1804.

[88] Haider, S., Parkinson, G.N., Neidle. S. Molecular Dynamics and Principal Components Analysis of Human Telomeric Quadruplex Multimers; *Biophys J* **2008**, 95, 296–311.

[89] Petraccone, L., Garbett, N.C., Chaires, J.B., Trent, J.O. An Integrated Molecular Dynamics (MD) and Experimental Study of Higher Order Human Telomeric Quadruplexes; *Biopolymers* **2010**, 93, 533–548.

[90] Cavallari, M., Calzolari, A., Garbesi, A., Di Felice, R. Stability and Migration of Metal Ions in G4-Wires by Molecular Dynamics Simulations; *J Phys Chem B* **2006**, 110, 26337–26348.

[91] Reshetnikov, R., Golovin, A., Spiridonova, V., Kopylov, A., Šponer, J. Structural Dynamics of Thrombin-Binding DNA Aptamer d(GGTTGGTGTGGTTGG) Quadruplex DNA Studied by Large-Scale Explicit Solvent Simulations; *J Chem Theory Comput* **2010**, 6, 3003–3014.

[92] Akhshi, P., Mosey, N.J., Wu, G. Free-Energy Landscapes of Ion Movement Through a G-Quadruplex DNA Channel; *Angew Chem* **2012**, 12, 2850-2854.

[93] Haider, S.M. and Neidle, S. A Molecular Model for Drug Binding to Tandem Repeats of Telomeric G-quadruplexes; *Biochem Soc Trans* **2009**, 37, 583–588.

[94] Hou, J.Q., Chen, S.B., Tan, J.H., Ou, T.M., Luo, H.B., Li, D., Xu, J., Gu, L.J., Huang, Z.S. New Insights into the Structures of Ligand−Quadruplex Complexes from Molecular Dynamics Simulations; *J Phys Chem B* **2010,** 114, 15301–15310.

[95] Collie, G.W, Haider, S.M, Neidle, S., Parkinson, G.N. A Crystallographic and Modelling Study of a Human Telomeric RNA (TERRA) Quadruplex; *Nucleic Acids Res* **2010**, 38, 5569–5580.

[96] Li, Q., Xiang, J., Li, X., Chen, L., Xu, X., Tang, L., Zhou, Q. *et al.* Stabilizing Parallel G-quadruplex DNA by a New Class of Ligands: Two Non-planar Alkaloids Through Interaction in Lateral Grooves; *Biochimie* **2009**, 91, 811–819.

[97] Chen, S.B., Tan, J.H., Ou, T.M, Huang, S.L., An, L.K., Luo, H.B., Li, D., Gu, L.Q., Huang, Z.S. Pharmacophore-based Discovery of Triaryl-substituted Imidazole as New Telomeric G-quadruplex Ligand; *Bioorg Med Chem Letters* **2011**, 21, 1004–1009.

[98] Ma, D.L., Pui-Yan Ma, V., Chan, D.S., Leung, K.H., Zhong, K.H, Leung, C.H. In Silico Screening of Quadruplex-binding Ligands; *Methods* **2012**, 57, 106-114.

[99] Neidle, S. Design Principles for Quadruplex-binding Small Molecules in *Therapeutic Applications of Quadruplex Nucleic Acids*; ed. Neidle, S; Academic Press Boston **2012**, 151–174.

[100] Totrov, M. and Abagyan, R. Flexible Protein-ligand Docking by Global Energy Optimization in Internal Coordinates; *Proteins* **1997**, 1, 215–220.

[101] Ma, D.L., Lai, T.S., Chan, F.Y., Chung, W.H., Abagyan, R., Leung, Y.C., Wong, K.Y. Discovery of a Drug−Like G−Quadruplex Binding Ligand by High−Throughput Docking; *ChemMedChem* **2008**, 3, 881–884.

[102] Lee, H.M, Chan, D.S., Yang, F., Lam, H.Y, Yan, S.C., Che, C.M., Ma, D.L., Leung, C.H. Identification of Natural Product Fonsecin B as a Stabilizing Ligand of C-myc G-quadruplex DNA by High-throughput Virtual Screening; *Chem Comm* **2010**, 46, 4680–4682.

[103] Ma, D.L., Chan, D.S., Leung, C.H. Molecular Docking for Virtual Screening of Natural Product Databases; *Chem Sci* **2011**, 2, 1656–1665.

[104] Cosconati, S., Marinelli, L., Trotta, R., Virno, A., Mayol, L., Novellino, E., Olson, A.J., Randazzo, A. Tandem Application of Virtual Screening and NMR Experiments in the Discovery of Brand New DNA Quadruplex Groove Binders; *J Am Chem Soc* **2009**, 131, 16336–16337.

[105] Trotta, R., De Tito, S., Lauri, I., La Pietra, V., Marinelli, L., Cosconati, S., Martino, L. *et al*. A More Detailed Picture of the Interactions Between Virtual Screening-derived Hits and the DNA G-quadruplex: NMR, Molecular Modelling and ITC Studies; *Biochimie* **2011**, 93, 1280–1287.

[106] Holt, P.A., Buscaglia, R., Trent, J.O., Chaires, J.B. A Discovery Funnel for Nucleic Acid Binding Drug Candidates; *Drug Dev Res* **2011**, 72, 178–186.

[107] Holt, P.A., Chaires, J.B., Trent, J.O. Molecular Docking of Intercalators and Groove-binders to Nucleic Acids Using Autodock and Surflex; *J Chem Info Model* **2008**, 48, 1602–1615.

[108] Holt, P.A., Ragazzon, P., Strekowski, L., Chaires, J.B., Trent, J.O. 2009. Discovery of Novel Triple Helical DNA Intercalators by an Integrated Virtual and Actual Screening Platform; *Nucleic Acids Res* **2009**, 37, 1280–1287.

[109] Phan, A.T., Kuryavyi, V., Burge, S., Neidle, S., Patel, J.D. Structure of an Unprecedented G-quadruplex Scaffold in the Human C-kit Promoter; *J Am Chem Soc* **2007**, 129, 4386–4392.

[110] Wei, D., Parkinson, G.N., Reszka, A.P., Neidle, S. Crystal Structure of a C-kit Promoter Quadruplex Reveals the Structural Role of Metal Ions and Water Molecules in Maintaining Loop Conformation; *Nucleic Acids Res* **2012**, 40, 4691-4700.

[111] Schlick, T. Biomolecular Structure and Modeling: Historical Perspective in *Molecular Modeling and Simulation: An Interdisciplinary Guide*; ed. Schlick, T.; Springer New York **2010**, 26, 1-40.

[112] Schlick, T. Theoretical and Computational Approaches to Biomolecular Structure in *Molecular Modeling and Simulation: An Interdisciplinary Guide*; ed. Schlick, T.; Springer New York **2010**, 26, 237-264.

[113] Barril, X. and Soliva, R. Molecular Modelling; *Mol BioSyst* **2006**, 2, 660–681.

[114] van Gunsteren, W.F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P. *et al* Biomolecular Modeling: Goals, Problems, Perspectives; *Angew Chem* **2006**, 45, 4064–4092.

[115] Jorgensen, W.L. The Many Roles of Computation in Drug Discovery; *Science* **2004**, 303, 1813–1818.

[116] Finn, P.W. and Kavraki, L.E. Computational Approaches to Drug Design; *Algorithmica* **1999**, 25, 347–371.

[117] van der Kamp, M.W., Shaw, K.E., Woods, C.J., Mulholland, A.J. Biomolecular Simulation and Modelling: Status, Progress and Prospects; *J R Soc Interface* **2008**, 5, S173-S190.

[118] Adcock, S.A. and McCammon, J.A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins; *Chem Rev* **2006**, 106, 1589–1615.

[119] Lipkowitz, K. Abuses of Molecular Mechanics: Pitfalls to Avoid; *J Chem Edu* **1995**, 72, 1070.

[120] Ditzler, M.A., Otyepka, M. Šponer, J., Walter, N.G. Molecular Dynamics and Quantum Mechanics of RNA: Conformational and Chemical Change We Can Believe In; *Acc Chem Res* **2010**, 43, 40–47.

[121] York, D.M., Darden, T.A., Pedersen, L.G. The Effect of Long−range Electrostatic Interactions in Simulations of Macromolecular Crystals: A Comparison of the Ewald and Truncated List Methods; *J Chem Phys* **1993**, 99, 8345–8348.

[122] Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., Pedersen, L.G. A Smooth Particle Mesh Ewald Method; *J Chem Phys* **1995**, 103, 8577.

[123] Dong, F., Olsen, B., Baker, N.A. Computational Methods for Biomolecular Electrostatics; *Methods Cell Biol* **2008**, 84, 843-870.

[124] Schlick, T. Molecular Dynamics: Basics in *Molecular Modeling and Simulation: An Interdisciplinary Guide*; ed. Schlick, T.; Springer New York **2010**, 26, 425-459.

[125] Ryckaert, J.P., Ciccotti, G., Berendsen, H.J. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of N-alkanes; *J Comp Phys* **1977**, 23, 327–341.

[126] Hess, B., Bekker, H., Berendsen, H.J., Fraaije, J.G. LINCS: A Linear Constraint Solver for Molecular Simulations; *J Comp Chem* **1997**, 18, 1463–1472.

[127] Numata, J. Entropy and Thermodynamics in Biomolecular Simulation in *Handbook of Research on Systems Biology Applications in Medicine*; ed. Daskalaki, A.; IGI Global **2008**, 731-758.

[128] Christ, C.D., Mark, A.E., van Gunsteren, W.F. Basic Ingredients of Free Energy Calculations: A Review; *J Comp Chem* **2010**, 31, 1569–1582.

[129] Jarzynski, C. Nonequilibrium Equality for Free Energy Differences; *Phys Rev Lett* **1997**, 78, 2690.

[130] Aqvist, J., Medina, C., Samuelsson, J.E. A new method for predicting binding affinity in computer-aided drug design; *Protein Eng* **1994**, 7, 385-391.

[131] Srinivasan, J., Cheatham, T.E., Cieplak, P., Kollman, P.A., Case, D.A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices; *J Am Chem Soc* **1998**, 120, 9401-9409.

[132] Qiu, D., Shenkin, P.S., Hollinger, F.P., Still, W.C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii; *J Phys Chem A* **1997**, 101, 3005-3014.

[133] Harris, S.A. and Laughton, C.A. A simple physical description of DNA dynamics: quasi-harmonic analysis as a route to the configurational entropy; *J Phys Condens Matter* **2007**, 19, 76103-76117.

[134] Rastelli, G., del Rio, A., Degliesposti, G., Sgobba, M. Fast and Accurate predictions of Binding Free Energies Using MM-PBSA and MM-GBSA; *J Comp Chem* **2010**, 31, 797-810.

[135] Bradshaw, R.T., Patel, B.H., Tate, E.W., Leatherbarrow, R.J., Gould, I.R. Comparing experimental and computational alanine scanning techniques for probing a prototypical protein–protein interaction; *Prot Eng Des Sel* **2011**, 24, 197-207.

[136] Hou, T., Wang, J., Li, Y., Wang, W. Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations; *J Chem Inf Model* **2011**, 51, 69–82.

[137] Kitchen, D., Decornez, B.H., Furr, J.R., Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications; *Nat Rev Drug Discov* **2004**, 3, 935–949.

[138] Klebe, G. Virtual Ligand Screening: Strategies, Perspectives and Limitations; *Drug Discov Today* **2006**, 11, 580–594.

[139] Jahn, A., Hinselmann, G., Fechner, N., Zell, A. Optimal Assignment Methods for Ligand-based Virtual Screening; *J Cheminform* **2009**, 1, 14.

[140] Halperin, I., Ma, B., Wolfson, H., Nussinov, R. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions; *Proteins* **2002**, 47, 409–443.

[141] Totrov, M. and Abagyan, R. Flexible Ligand Docking to Multiple Receptor Conformations: a Practical Alternative; *Curr Opin Struct Biol* **2008,** 18, 178-184.

[142] Neves, M.A., Totrov, M., Abagyan, R. Docking and Scoring with ICM: The Benchmarking Results and Strategies for Improvement; *J Comput Aided Mol Des* **2012**, 26, 675–686.

[143] Schneider, G. Virtual Screening: An Endless Staircase?; *Nat Rev Drug Discov* **2010**, 9, 273–276.

[144] Scior, T., Bender, A., Tresadern, G., Medina-Franco, J.L., Martínez-Mayorga, K., Langer, T., Cuanalo-Contreras, K., Agrafiotis, D.K. Recognizing Pitfalls in Virtual Screening: A Critical Review; *J Chem Info Model* **2012**, 52, 867-881.

[145] Ruvinsky, A.M. Role of binding entropy in the refinement of protein-ligand docking predictions: analysis based on the use of 11 scoring functions; *J Comput Chem* **2007**, 28, 1364–1372.

[146] Takahashi, O., Kohno, Y., Nishio, M. Relevance of weak hydrogen bonds in the conformation of organic compounds and bioconjugates: evidence from recent experimental data and high-level ab initio MO calculations; *Chem Rev* **2010**, 110, 6049–6076.

[147] Roberts, B.C. and Mancera, R.L. Ligand-protein docking with water molecules; *J Chem Inf Model* **2008**, 48, 397–408.

[148] Sim, A.Y., Minary, P. and Levitt, M. Modeling Nucleic Acids; *Curr Opin Struct Biol* **2012**, 22, 273-278.

[149] Ricci, C.G., de Andrade, A.S., Mottin, M., and Netz, P.A., Molecular Dynamics of DNA: Comparison of Force Fields and Terminal Nucleotide Definitions; *J Phys Chem B* **2010**, 114, 9882–9893.

[150] Mackerell, A.D., Jr., Empirical Force Fields for Biological Macromolecules: Overview and Issues; *J Comp Chem* **2004**, 25, 1584–1604.

[151] Harris, S.A. Modelling the Biomechanical Properties of DNA Using Computer Simulation; *Phil Trans R Soc A* **2006**, 364, 3319-3334.

[152] Laughton, C.A. Molecular Dynamics Simulations of DNA Triple-Helices. Does The Triplex d(A)10·d(T)10·d(T)10 Have A-Form or B-form Geometry?; *Molecular Simulation* **1995**, 14, 275-289.

[153] Mitchell, J.S., Laughton, C.A., Harris, S.A. Atomistic simulations reveal bubbles, kinks and wrinkles in supercoiled DNA; *Nucleic Acids Res* **2011**, 39, 3928-3938.

[154] Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J. GROMACS: fast, flexible, and free; *J Comput Chem* **2005**, 26, 1701-1718.

[155] Hess, B., Kutzner, C., Van der Spoel, D., Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation; *J Chem Theory Comput* **2008**, 4, 435−447.

[156] Schuler, L.D., Daura, van Gunsteren, W.X. An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J Comp Chem* **2001**, 22, 1205-1218.

[157] Jorgensen, W.L. and Tirado-Rives, J. The OPLS (optimized potentials for liquid simulations) potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin; *J Am Chem Soc* **1988**, 110, 1657-1666.

[158] Weiner, P.K. and Kollman, P.A. AMBER: Assisted Model Building with Energy Refinement. A General Program for Modeling Molecules and Their Interactions; *J Comp Chem* **1981**, 2, 287-303.

[159] Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan,S., Karplus, M. CHARMM: A program for macromolecular energy, minmimization, and dynamics calculations; *J Comp Chem* **1983**, 4, 187-217.

[160] Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J. The Amber Biomolecular Simulation Programs; *J Comp Chem* **2005**, 26, 1668–1688.

[161] Perez, A., Marchan, I., Svozil, D., Šponer, J., Cheatham, T.E., Laughton, C.A., Orozco, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of Alpha/gamma Conformers; *Biophys J* **2007**, 92, 3817–3829.

[162] Varnai, P. and Zakrzewska, K. DNA and its counterions: a molecular dynamics study; *Nucleic Acids Res* **2004**, 32, 4269–4280.

[163] Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., Kollman, P.A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules; *J Am Chem Soc* **1995**, 117, 5179–5197.

[164] Sorin, E.J. and Pande, V.S. Exploring the Helix-Coil Transition via All-Atom Equilibrium nsemble Simulations. *Biophys J* **2005**, 88, 2472-2493.

[165] Homeyer, N., Horn, A.H., Lanig, H., Sticht, H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine,

phosphothreonine, phosphotyrosine, and phosphohistidine; *J Mol Model* **2006**, 12, 281-289.

[166] Lindorff-Larsen, K. Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., Shaw, D.E. Improved side-chain torsion potentials for the Amber ff99SB protein force field; *Proteins* **2010**, 78, 1950-1958.

[167] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., Simmerling, C. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters; *Proteins* **2006**, 65, 712-725.

[168] Ponder, J.W. and Case, D.A. Force Fields for Protein Simulations; *Adv Protein Chem* **2003**, 66, 27–85.

[169] Sousa da Silva, A.W., Vranken, W.F. ACPYPE - AnteChamber PYthon Parser interfacE; *BMC Res Notes* **2012**, 23, 367.

[170] Wang, J., Wang, W., Kollman, P.A., Case, D.A. Automatic atom type and bond type perception in molecular mechanical calculations; *J Mol Graphics Model* **2006**, 25, 247–260.

[171] Ryckaert, J.P. and Bellemans, A. Molecular Dynamics of Liquid Alkanes; *Faraday Discuss Chem Soc* **1978**, 66, 95-106.

[172] Cohen, P. The origins of protein phosphorylation; *Nat Cell Biol* **2002**, 4, E127- E130.

[173] Wojciechowski, M.T., Grycuk, J.M., Antosiewicz, Lesyng, M. Prediction of Secondary Ionization of the Phosphate Group in Phosphotyrosine Peptides; *Biophys J* **2003**, 84, 750–756.

[174] Dourlat, J., Valentin, B., Liu, W.Q., Garbay, C. New Syntheses of Tetrazolylmethylphenylalanine and O-malonyltyrosine as pTyr Mimetics for the Design of STAT3 Dimerization Inhibitors; *Bioorg Med Chem Letters* **2007**, 17, 3943–3946.

[175] Holland, S.M., DeLeo, F.R., Elloumi, H.Z., Hsu, A.P., Uzel, G., Brodsky, N., Freeman, A.F., Demidiwich, A., Davis, J., Turner, M.L., Anderson, V.L., Darnell, D.N., Welch, P.A., Kuhns, D.B, Frucht, D.M., Malech, H.M., Gallin, J.I., Kobayashi, S.D., Whitney, A.R., Voyich, J.M., Musser, J.M., Woellner, C., Schäffer, A.A., Puck, J.M., Grimbacher, B. STAT3 Mutations in the Hyper-IgE Syndrome; *N Engl J Med* **2007**, 357, 1608-1619.

[176] Heimall, J., Freeman, A., Holland, S.M. Pathogenesis of Hyper IgE Syndrome; *Clinic Rev Allerg Immunol* **2010**, 32-38.

[177] Darnell, J.E. Validatinng Stat3 in cancer therapy; *Nat Med* **2005**, 11, 595-596.

[178] Schröder, M., Kroeger, K.M., Volk, H.D., Eidene, K.A., Grütz, G. Preassociation of nonactivated STAT3 molecules demonstrated in living cells using bioluminescence resonance energy transfer: a new model of STAT activation?; *J Leukoc Biol* **2004**, 75, 792-797.

[179] Timofeeva, O.A., Chasovskikh, S., Lonskaya, I., Tarasova, N.I., Khavrutskii, L., Tarasov, S.G., Zhang, X., Korostyshevskiy, V.R., Cheema A., Zhang, L., Dakshanamurthy, S., Brown,M.L., Dritschilo, A. Mechanisms of unphosphorylated STAT3 transcription factor binding to DNA; *J Biol Chem* 2012, **287**, 14192–14200.

[180] Thurston, D.E. and Zinzalla, G. Targeting protein-protein interactions for therapeutic intervention: a challenge for the future; *Future Med Chem* **2009**, 1, 65-93.

[181] Becker, S., Groner, B., Muller, C.W. Three-dimensional structure of the STAT3 [beta] homodimer bound to DNA; *Nature* **1998**, 394, 145−151.

[182] Ren, Z., Mao, X., Mertens, C., Krishnaraj, R., Qin, J., Mandal, P. K., Romanwski, M.J., McMurray, J.S., Chen, X. Crystal structure of unphosphorylated STAT3 core fragment. Biochem Biophys Res Commun **2008**, 374, 1−5.

[183] Lin, J., Buettner, R., Yuan, Y.C., Yip, R., Horne, G., Jove, R., Vaidehi, N. Molecular dynamics simulations of the conformational changes in signal transducers and activators of transcription, STAT1 and STAT3; *J Mol Graphics Modell* **2009**, 28, 347−356.

[184] Park, I.-H., Li, C. Characterization of molecular recognition of STAT3 SH2 domain inhibitors through molecular simulation; *J Mol Recognit* **2010**, 23, 1−12.

[185] Baron, R., Setny, P., McCammon, A.J. Water in cavity− ligand recognition; *J Am Chem Soc* **2010**, 132, 12091−12097.

[186] Henchman, R.H., McCammon, J.A. Structural and dynamic properties of water around acetylcholinesterase; *Protein Sci* **2002**, 11, 2080−2090.

[187] De Simone, A., Dodson, G.G., Verma, C.S., Zagari, A., Fraternali, F. Prion and water: tight and dynamical hydration sites have a key role in structural stability; *Proc Natl Acad Sci USA* **2005**, 102, 7535−7540.

[188] Scorciapino, M.A., Robertazzi, A., Casu, M., Ruggerone, P., Ceccarelli, M. Heme proteins: the role of solvent in the dynamics of gates and portals; *J Am Chem Soc* **2010**, 132, 5156−63.

[189] Reddy, C.K., Das, A., Jayaram, B. Do water molecules mediate protein-DNA recognition?; *J Mol Biol* **2001**, 314, 619−632.

[190] Jayaram, B., Jain, T. The role of water in protein-DNA recognition; *Annu Rev Biophys Biomol Struct* **2004**, 33, 343−361.

[191] Minegishi, Y., Saito, M., Tsuchiya, S., Tsuge, I., Takada, H., Hara, T., Kawamura, N., Ariga, T., Pasic, S., Stojkovic, O., Metin, A., Karasuyama, H. Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome; *Nature* **2007**, 448, 1058−1062.

[192] Woellner, C. *et al.* Mutations in STAT3 and diagnostic guidelines for hyper-IgE syndrome; *J Allergy Clin Immunol* **2010**, 125, 424−432.

[193] Cole, C., Barber, J.D., Barton, G.J. The JPRED3 secondary structure prediction server; *Nucleic Acids Res* **2008**, 36, W197-W201.

[194] Fischer A., Sali, A. ModLoop: automated modeling of loops in protein structures; *Bioinformatics* **2003**, 19, 2500-2501.

[195] The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC. http://pymol.org/

[196] Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L. Comparison of simple potential functions for simulating liquid water; *J Chem Phys* **1983**, 79, 926-935.

[197] Lin, Y. and Lim, C. Factors Governing the Protonation State of Zn-Bound Histidine in Proteins: A DFT/CDM Study; *J Am Chem Soc* **2002**, 126, 2602-2612.

[198] Li, H., Robertson, A.D., Jensen, J.H. Very Fast Empirical Prediction and Interpretation of Protein pKa Values; *Proteins* **2005**, 61, 704-721.

[199] Bussi, G., Donadio, D., Parrinello, M. Canonical sampling through velocity rescaling; *J Chem Phys* **2007**, 126, 014101.
[200] Parinello, M. Polymorphic transition in single crystals: A new molecular dynamics method; *J Appl Phys* **1981**, 52, 7182-7189.
[201] Nose, S. and Klain, M.L. Constant pressure molecular dynamics for molecular systems; *Mol Phys* **1983**, 50, 1055-1076.
[202] Humphrey, W., Dalke, A., Schulten, K. VMD: Visual molecular dynamics; *J Mol Graphics* **1996**, 14, 33−38. (http://www.ks.uiuc.edu/Research/vmd/)
[203] Amadei, A., Linssen, A. B., Berendsen, H. J. Essential dynamics of proteins; *Proteins* **1993**, 17, 412−425.
[204] Mesentean, S., Fischer, S., Smith, J.C. Analyzing large-scale structural change in proteins: comparison of principal component projection and Sammon mapping; *Proteins* **2006**, 64, 210−218.
[205] Tai, K., Shen, T., Börjesson, U., Philippopoulos, M., McCammon, J.A. Analysis of a 10-ns molecular dynamics simulation of mouse acetylcholinesterase; *Biophys J* **2001**, 81, 715−724.
[206] Bock, H.H. Probabilistic models in cluster analysis; *Comp Stat Data Anal* **1996**, 23, 5−28.
[207] Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W.F., Mark, A.E. Peptide folding: when simulation meets experiment; *Angew Chem* **1999**, 38, 236−240.
[208] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Ferrin, T.E. UCSF Chimera - a visualization system for exploratory research and analysis; *J Comput Chem* **2004**, 25, 1605− 1612.
[209] Kleywegt, G.J. and Jones, T.A. Where freedom is given, liberties are taken; *Structure* **1995**, 3, 535-540. http://xray.bmc.uu.se/usf/dejavu.html
[210] The CCP4 suite: programs for protein crystallography; *Acta Crystallogr D Biol Crystallogr* **1994**, 50, 760-763. http://www.ccp4.ac.uk/
[211] Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., Honig, B. The role of DNA shape in protein-DNA recognition; *Nature* **2009**, 461, 1248−1253.
[212] Patrick, C. Drugs targeting protein-protein interactions; *ChemMedChem* **2006**, 1, 400-411.
[213] Drewry, J.A., Burger, S., Mazouchi, A. Duodu, E., Ayers, P., Gradinaru, C.C., Gunning, P.T. Src Homology 2 Domain Proteomimetics: Developing Phosphopeptide Selective Receptors; *MedChemComm* **2012**, 3, 763–770.
[214] Filippakopoulos, P., Müller, S., Knapp. S. SH2 Domains: Modulators of Nonreceptor Tyrosine Kinase Activity; *Curr Opin Struct Biol* **2009**, 19, 643–649.
[215] McMurray, J.S. Structural Basis for the Binding of High Affinity Phosphopeptides to Stat3; *Biopolymers* **2008**, 90, 69–79.
[216] Haftchenary, S., Avadisian, M. Gunning, P.T. Inhibiting Aberrant Stat3 Function with Molecular Therapeutics; *Anti-Cancer Drugs* **2011**, 22, 115–127.
[217] Watanabe, K., Saito, K., Kinjo, M., Matsuda, M., Tamura, M., Kon, S., Miyazaki, T., Uede, T. Molecular Dynamics of STAT3 on IL-6 Signaling Pathway in Living Cells; *Biochem Biophys Res Commun* **2004**, 324, 1264–1273.
[218] Mohr, A. Chatain, N., Domoszlai, T., Rinis, N., Sommerauer, M., Vogt, M., Müller-Newen, G. Dynamics and Non-canonical Aspects of JAK/STAT Signalling; *Eur J Cell Biol* **2012**, 91, 524-532.

[219] Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E.,Villa, E., Chipot, C., Skeel, R., Kale, L., Schulten, K. Scalable molecular dynamics with NAMD; *J Comp Chem* **2005**, 26, 1781-1802.

[220] Husby, J. Todd, A.K., Haider, S.M., Zinzalla, G., Thurston, D.E., Neidle, S. Molecular dynamics studies of the STAT3 homodimer:DNA complex: relationships between STAT3 mutations and protein-DNA recognition; *J Chem Inf Model* **2012**, 52, 1179-1192.

[221] Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J. Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity; *J Comp Chem* **2009**, 16: 2785-91.

[222] Dhanik, A., McMurray, J.S., Kavraki, L. On Modeling Peptidomimetics in Complex with the SH2 Domain of Stat3; in *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2011,** 3229–3232.

[223] Jones, G., Willett, P., Glen, R.C, Leach, A.R., Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking; *J Mol Biol* **1997**, 267, 727–748.

[224] Lang, T.P., Brozell, S.R., Mukherjee, S., Pettersen, E.F., Meng, E.C., Thomas, V., Rizzo, R.C., Case, D.A., James, T.L., Kuntz, I.D. DOCK 6: Combining Techniques to Model RNA–small Molecule Complexes; *RNA* **2009**, 15, 1219–1230.

[225] Jakalian, A., Jack, D.B., Bayly, C.I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation; *J Comput Chem* **2002**, 23, 1623-1641.

[226] Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P., Case, D.A. Development and Testing of a General Amber Force Field; *J Comput Chem* **2004**, 25, 1157-1174.

[227] Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E. A geometric approach to macromolecule-ligand interactions; *J Mol Biol* **1982**, 161, 269-288.

[228] Richards, F M. Areas, Volumes, Packing and Protein Structure; *Annu Rev Biophys Bioeng* **1977**, 6, 151–176.

[229] Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, B., Chong, L., Lee, M. Lee, T., Duan, Y., Wang, W. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models; *Acc Chem Res* **2000**, 33, 889–897.

[230] Hawkins, G.D., Cramer, S.J., Truhlar, D.G. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium; J Phys Chem **1996**, 100, 19824–19839.

[231] Tsui, V. and Case, D.A. Theory and Applications of the Generalized Born Solvation Model in Macromolecular Simulations; *Biopolymers* **2000**, 56, 275–291.

[232] Liebeschuetz, J.W., Cole, J.C., Korb, O. Pose Prediction and Virtual Screening Performance of GOLD Scoring Functions in a Standardized Test.; *J Comput Aided Mol Des* **2012**, 26, 737–748.

[233] Sousa, S.F., Fernandes, P.A., Ramos, J.M. Protein-ligand Docking: Current Status and Future Challenges; *Proteins* **2006**, 65, 15–26.

[234] Onufriev, A., Bashford, D. Case, D.A. Exploring Protein Native States and Large-scale Conformational Changes with a Modified Generalized Born Model; *Proteins* **2004**, 55, 383–394.

[235] Sanders, M.P., Barbosa, A.J., Zarzycka, B., Nicolaes, G.A., Klomp, J.P., de Vlieg, J., Del Rio, A. Comparative Analysis of Pharmacophore Screening Tools; *J Chem Info Model* **2012**, 52, 1607–1620.

[236] Wolber, G., Dornhofer, A., Langer, T. Efficient Overlay of Small Organic Molecules Using 3D Pharmacophores; *J Comput Aided Mol Des* **2006**, 20, 773–788.

[237] Wolber, G. and Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters; *J Chem Info Model* **2005**, 45, 160–169.

[238] Haan, S., Hemmann, U., Hassiepent, U., Schapert, F., Schneider-Mergeners, J., Wollmer, A., Heinrich, P.C., Grötzinger, J. Characterization and Binding Specificity of the Monomeric STAT3-SH2 Domain; *J Biol Chem* **1999**, 274, 1342–1348.

[239] McMurray, J.S. A New Small-Molecule Stat3 Inhibitor; *Chemistry & Biology* **2006**,13, 1123–1124.

[240] Gohlke, H. and Case, D.A. Converging Free Energy Estimates: MM-PB(GB)SA Studies on the Protein-protein Complex Ras-Raf; *J Comp Chem* **2004**, 25, 238–250.

[241] Mandal, P.K., Limbrick, D., Coleman, D., Dyer, G.A., Ren, Z., Sanderson Birtwistle, J., Xiong, C., Chen, X., Briggs, J.M., McMurray, J. Conformationally Constrained Peptidomimetic Inhibitors of Signal Transducer and Activator of Transcription. 3: Evaluation and Molecular Modeling; *J Med Chem* **2009**, 52 2429–2442.

[242] Williamson, J.R., Rarguramab, M.K., Cech, T.R. Monovalent cation-induced structure of telomeric DNA: The G-quartet model; *Cell* **1989**, 59, 871-880.

[243] Sen, D. and Gilbert, W. A sodium-potassium switch in the formation of four-stranded G4-DNA; *Nature* **1990**, 344, 410-414.

[244] Henderson, E., Hardin, C.C., Walk, S.K., Tinoco, I. Jr., Blackburn, E.H. Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine·guanine base pairs; *Cell* **1987**, 51, 899-908.

[245] Huppert, J.L. and Balasubramanian, S. Prevalence of quadruplexes in the human genome; *Nucleic Acids Res* **2005**, 33, 2908-2916.

[246] Huppert, J.L. and Balasubramanian, S. G-quadruplexes in promoters throughout the human genome; *Nucleic Acids Res* **2007**, 35, 406-413.

[247] Neidle, S. The structures of quadruplex nucleic acids and their drug complexes; *Curr Opin Struct Biol* **2009**, 19, 239-250.

[248] Monchaud, D. and Teulade-Fichou, M.P. A hitchhiker's guide to G-quadruplex ligands; *Org Biomol Chem* **2008**, 6, 627-636.

[249] Shin-ya, K., Wierzba, K., Matsuo, K., Ohtani, T., Yamada, Y., Furihata, K. Hayakawa, Y., Seto, H. Telomestatin, a novel telomerase inhibitor from Streptomyces anulatus; *J Am Chem Soc* **2001**,123,1262–1263.

[250] Campbell, N.H., Collie, G.W., Neidle, S. Crystallography of DNA and RNA G-quadruplex nucleic acids and their ligand complexes; *Curr Protoc Nucleic Acid Chem* **2012**, 50, 17.6.1-17.6.22.

[251] Rodriguez, R., Müler, S., Yeoman, J.A., Trentesaux, C., Riou, J.F., Balasubramanian, S. A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. *J Am Chem Soc* **2008**, 130, 15758-15759.

[252] Rodriguez, R. *et al.* Small-molecule-induced DNA damage identifies alternative DNA structures in human genes; *Nat Chem Biol* **2012**, 8, 301–310.

[253] Cheng, M.K, Modi, C., Cookson, J.C., Hutchinson, I., Heald, R.A., McCarroll, A.J., Missailidis, S., Tanious, F., Wilson, D., Mergny, J.L., Laughton, C.A., Stevens, M.F. Antitumor Polycyclic Acridines. 20. Search for DNA Quadruplex Binding Selectivity in a Series of 8,13-dimethylquino[4,3,2-kl]acridinium Salts: Telomere-targeted Agents; *J Med Chem* **2008***,* 51, 963–975.

[254] Gavathiotis, E., Heald, R. A., Stevens, M. F., Searle, M. S. Drug recognition and stabilisation of the parallel-stranded DNA quadruplex d(TTAGGGT)4 containing the human telomeric repeat; *J Mol Biol* **2003**, 334, 25–36.

[255] Oliphant, T. E. Python for Scientific Computing; *Computing Science Engg* **2007**, 9, 10-20.

[256] Stewart, J.J. Application of the PM6 method to modeling proteins; *J Mol Model* **2009**, 15, 765–805.

[257] Labute, P. The Generalized Born/volume Integral Implicit Solvent Model: Estimation of the Free Energy of Hydration Using London Dispersion Instead of Atomic Surface Area; *J Comp Chem* **2008**, 29, 1693–1698.

[258] Řezáč, J., Fanfrlik, J., Salahub, D., Hobza, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes; *J Chem Theory Comput* **2009**, 5, 1749–1760.

[259] Klamt, A. and Schuurmann, G. COSMO: a New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient; *J Chem Soc Perkin Trans* 1993, 2, 799.

[260] The R Project for Statistical Computing; http://www.r-project.org

[261] Hurley, L.H., Von Hoff, D.D., Siddiqui-Jain, A., Yang, D. Drug Targeting of the c-MYC Promoter to Repress Gene Expression via a G-Quadruplex Silencer Element; *Seminars in Oncology* **2006**, 33, 498–512.

[262] Liu, J.-N. *et al.* Inhibition of myc promoter and telomerase activity and induction of delayed apoptosis by SYUIQ-5, a novel G-quadruplex interactive agent in leukemia cells. *Leukemia* **21**, 1300–1302 (2007).

[263] Gunaratnam, M., Swank, S., Haider, S. Galesa, K.,Rezska, A.P., Beltran, M., Cuenca, F., Fletcher, J.A., Neidle, S. Targeting human gastrointestinal stromal tumor cells with a quadruplex-binding small molecule; *J Med Chem* **2009**, 52, 3774–3783.

[264] Balasubramanian, S. and Neidle, S. G-quadruplex nucleic acids as therapeutic targets; *Curr Opin Chem Biol* **2009**, 13, 345–353.

[265] Drygin, D., Siddiqui-Jain, A., O'Brien, S., Schwaebe, M., Lin, A., Bliesath, J., Ho,C.B., Proffitt, C., Trent, K., Whitten, J., Lim, J., Van Hoff, D., Anderes, K., Rice, W.G. Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis; *Cancer Res* **2009**, 69, 7653–7661.

[266] Yang, W., Fucini, R.W., Fahr, B., Randal, M., Lind, K.E., Lam, M., Lu, W., Lu, Y., Cary, D., Romanowski, M.J. Fragment-Based Discovery of Nonpeptidic BACE-1 Inhibitors Using Tethering; *Biochemistry* **2009**, 48, 4488–4496.

[267] Deigan, K. E. and Ferré-D'Amaré, A.R. Riboswitches: discovery of drugs that target bacterial gene-regulatory RNAs; *Acc Chem Res* **2011**,44, 1329–1338.

[268] Ambrus, A., Chen, D., Dai, J., Jones, R.A., Yang, D. Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization; *Biochemistry* **2005**, 15, 2048-2058.

[269] Allinger, N.L. Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms; *J Am Chem Soc* **1977**, 99, 8127-8134.

[270] Borhani, D.W. and Shaw, D.E. The future of molecular dynamics simulations in drug discovery; *J Comput Aided Mol Des* **2012**, 26, 15-26.

APPENDIX

PUBLICATIONS